UNIVERSITY OF CALIFORNIA
RIVERSIDE

An ‘Omics Investigation into the Effects of Aluminum Toxicity on Arabidopsis *Thaliana*

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics, and Bioinformatics

by

Stephen Christopher Bolaris

March 2019

Dissertation Committee:
Dr. Jason Stajich, Chairperson
Dr. Connie Nugent
Dr. Thomas Girke

The Dissertation of Stephen Christopher Bolaris is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements:

Thank you to my committee for supporting me and helping to complete this work.

To Thomas Alessandri:

The man who taught me that nothing is impossible,

somethings are just more difficult than others

RIP

To my friends and family for the support and encouragement while in grad school

Funding for this project was provided by the National Science Foundation (NSF) and European Research Area Network for Coordinating Action in Plant Sciences (ERA-CAPS) provided through the Larsen Lab at UC Riverside.

ABSTRACT OF THE DISSERTATION

An ‘Omics Investigation into the Effects of Aluminum Toxicity on Arabidopsis *Thaliana*

by

Stephen Christopher Bolaris

Doctor of Philosophy, Graduate Program in Genetics, Genomics, and Bioinformatics
University of California, Riverside, March 2019
Dr. Jason Stajich, Chairperson

Aluminum (Al) toxicity is a global problem that leads to stoppage of root growth, overall smaller plant size and lower crop yields. Previous research has shown the molecular response of plants to Al toxicity occurs through a DNA damage response pathway involving ATR and SOG1 genes. To explore this phenomenon further both transcriptomic and genomic experiments were performed using Arabidopsis *Thaliana*. The goal of the transcriptomics was to determine a gene or suite of genes that were deferentially expressed with Al^{3+} exposure that could potentially confer Al tolerance to crop plants. While a companion genomics study aimed to understand what type of genomic damage was occurring following Al exposure. Arabidopsis seedlings were grown on gel soaked media plates in the absences or presence of Al^{3+}, before nucleic acids were harvested for Illumina short read sequencing. Transcriptionally, a suite of genes that included known Al response factors and some novel genes were identified using a cut off of 2 fold change and a false discover rate of 1%, 10 of the genes had their

expression validated using quantitative real time PCR. In addition, it was identified genetically that Al toxicity leads to the generation of one and two base pair insertions and deletions, which were determined to be statistically significant. With this knowledge future experiments can be performed with the promise of finding the molecular citiical to responding to Al exposure and how to use this response to confer Al tolerance to crop plants. Such experiments should include testing Arabidopsis mutants that have reactive oxygen species related genes knocked out or overexpressed to evaluate the level of genomic damage in the presence Al^{3+}. Additionally, genes identified from this transcriptional study should have their expression modified to further understand their role in Al toxicity. Pathway interaction studies with these factors could highlight the full molecular pathway of the plants response to Al exposure.

Table of Contents

List of Figures

List of Tables

Introduction

A Global Problem

Greater than 30% of the world's arable land is comprised of acidic soils (pH<5.5) (Figure 1) [1]. Aluminum (Al), which is the most abundant metal in the Earth's crust, speciates in these acidic soils to its cationic and highly toxic form, Al^{3+}. Most plants in these regions have severely compromised growth due to Al toxicity. Soil acidification, which results in Al speciation, where by Al which is biologically inert can disassociate with other elements in the soil becoming Al^{3+}. Some Al deposition is natural and some are considered to be man-made. As humans continue to change their environment, this alters the natural balance of their surrounding ecosystem and leads to phenomena such as acid rain. Some other factors that can lead to change in pH include industrial runoff, planting of crops, and use of fertilizers. Agricultural productivity on these soils is limited because Al^{3+} toxicity leads to severe root growth inhibition and overall reduced plant size and crop yields.

This is important to note since most of the regions in which Al toxicity is problematic are also the areas of the world with developing countries that do not possess the same means of modifying soil as more developed countries. Agricultural lime is one of the primary ways in which soil pH is modified since the addition of lime will alkalinize the soil, yet this is expensive and may need to be done routinely to maintain higher pH in order to keep Al in its biologically inert state.

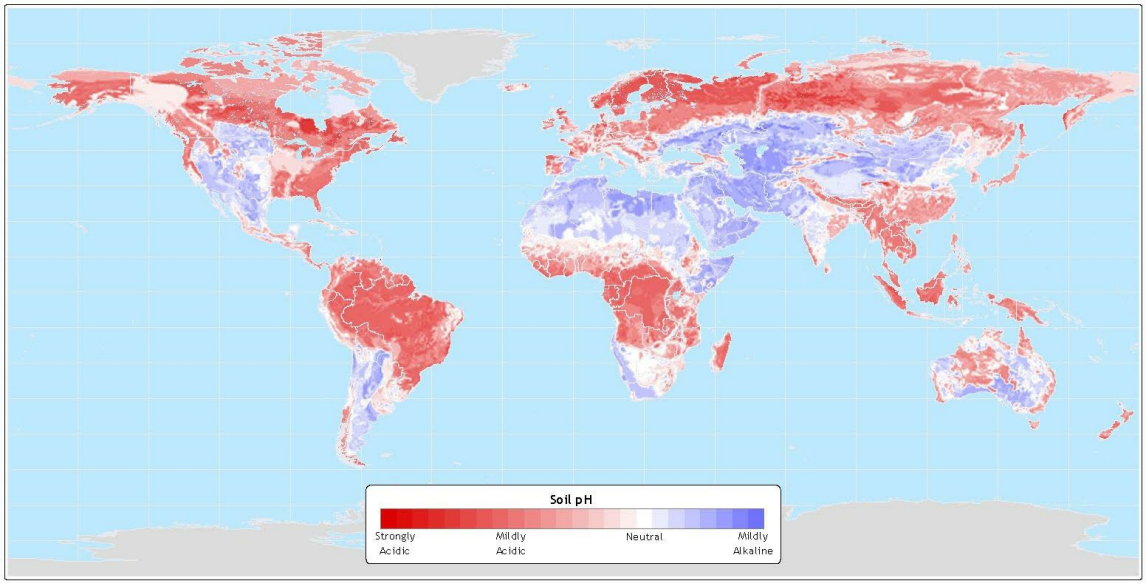


Figure 1: Global Map of Acidic Soils
Map of pH of the soils across the world with increasing red, lower in pH and blue, more basic. IGBP-DIS (1998) SoilData(V.0) A program for creating global soil-property databases, IGBP Global Soils Data Task, France.

Responses to Aluminum exposure

It should be noted that these highly Al toxic regions are some of the most biologically diverse areas in the world. Clearly, native plants have evolved to grow in these regions and have developed mechanisms to cope with Al in their soil environment. Three examples that have evolved mechanisms to cope with Al toxicity are hydrangea, buckwheat and tea plants. These types of plants are unique since they are considered to be Al accumulators, which occurs as a result of these plants having enhanced mechanisms of aluminum tolerance that allow them to accumulate millimolar levels of Al within their leaves as complexes with organic acids such as citrate [2].

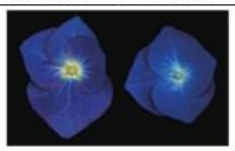
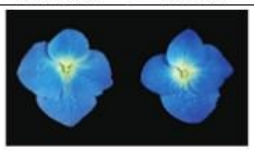
These Al accumulators possess unique genes that allow them to detoxify any internalized Al by using the function of H. macrophylla vacuolar aluminum transporter (hmVALT) and plasma membrane aluminum transporter (hmPALT) to sequester the Al in vesicles thus preventing it from interacting with more sensitive regions [3]. In sequestering Al in this way the plant also undergoes changes, and in the case of hydrangea these changes lead to the pigments of the petals changing through concentration dependent Al binding to aquaporins such as delphinidin-3- glucoside, and 3-caffeolylquinic acid which are located in the sepal cells [2]. The results can be seen in the plants through the changing color with a decrease in soil pH from blue to pink.

It has also been demonstrated that exclusion of Al from the root, or Al resistance, is an effective means to increase growth in Al toxic environments. Most commonly, Al resistant plants have increased release of organic acids such as malate or citrate into the rhizosphere. Exuded organic acids (OAs) complex with Al^{3+} to prevent it from being internalized and causing detrimental effects to the organism.

Table 2. Images, chromatic values, and pH of the sepals of *Hydrangea* grown in alkaline and acid soils.

	Ruby Red		HH9	
	Alkaline soil	Acid soil	Alkaline soil	Acid soil
Sepal color				
L*	50.0±3.7	37.2±3.1	39.6±2.7	36.6±1.4
a*	42.6±1.7	39.0±1.0	41.4±1.3	35.1±3.3
b*	−1.5±3.5	−8.9±1.1	−10.1±1.5	−15.1±1.8
Hue-angle (°)	358.0	347.2	346.3	336.8
pH	4.0±0.1	4.0±0.1	3.9±0.1	4.4±0.1

	HH13		HH19	
	Alkaline soil	Acid soil	Alkaline soil	Acid soil
Sepal color				
L*	46.3±2.3	44.1±2.3	60.6±2.3	46.6±1.3
a*	49.1±0.5	35.8±1.8	38.5±2.3	47.4±1.9
b*	−0.4±0.9	−14.1±1.4	−0.8±0.9	−4.6±1.0
Hue-angle (°)	359.5	338.5	358.9	354.4
pH	3.9±0.1	4.0±0.0	4.3±0.1	4.0±0.0

	Blue Sky		HH11		HH12	
	Alkaline soil	Acid soil	Alkaline soil	Acid soil	Alkaline soil	Acid soil
Sepal color						
L*	37.4±1.6	35.8±2.1	49.5±1.7	44.7±1.4	58.9±0.8	56.1±0.3
a*	25.6±2.8	24.9±1.5	11.5±0.7	15.2±1.7	9.3±0.1	11.4±0.7
b*	−38.3±1.0	−44.3±1.0	−35.1±0.3	−37.5±0.5	−28.5±0.8	−32.7±0.4
Hue-angle (°)	303.7	299.3	288.1	292.1	288.0	289.2
pH	4.0±0.1	4.1±0.1	4.2±0.0	4.3±0.1	4.5±0.1	4.3±0.1

	Frau Yoshiko		Frau Yoshimi		HH2	
	Alkaline soil	Acid soil	Alkaline soil	Acid soil	Alkaline soil	Acid soil
Sepal color						
L*	63.5±2.9	50.7±1.7	44.5±1.0	30.3±0.6	52.2±4.8	47.6±0.6
a*	36.0±3.3	22.7±1.0	46.4±1.2	28.0±0.1	37.5±2.8	27.0±2.0
b*	−5.2±0.6	−29.7±1.5	−2.3±0.8	−30.9±2.3	−10.2±0.6	−27.3±1.9
Hue-angle (°)	351.7	307.2	357.2	312.2	344.7	314.7
pH	4.1±0.1	4.3±0.1	4.1±0.1	4.1±0.1	4.1±0.1	4.1±0.1

Figure 2: Color Change of Hydrangea Petals
Table 2 taken from Kodama et al [4] demonstrating the color change in hydrangea in response to a change in soil pH and accumulation of Al^{3+}. With each line of hydrangea in alkaline and acidic soil.

Examples of this type of Al resistance mechanism can be seen in various crop plants including wheat, sorghum, maize, and rice. Three organic acids (OAs) that are key metabolic intermediates have been found to confer Al resistance when exuded at high levels including malate, citrate, oxalate. Citrate was originally found to be key to Al resistance in snap beans (*Phaseolus vulgaris* L. "Romano" and "Dade") [5] whereas malate is important to wheat (*Triticum aestivum* L.) [6]. These OAs are used to chelate Al^{3+} and while often there is a predominant OA that is released, some plants will release combinations [7,8]. While exudation of OAs by the root is an effective way to exclude Al, some crop species such as wheat also use these OAs to chelate the Al internally as a means to confer Al tolerance [9].

While OAs are thought of as the main mechanism of Al resistance, there are other ways a plant can alter the soil in an attempt to prevent Al^{3+} from being uptaken. One such way is the use of phenols [10], which due to their ring structure can chelate Al^{3+} although less effectively than OAs [11]. This has been studied in maize with certain phenols being more prevalent than others such as catechol, catechin, and quercetin, which tend to be released along with citrate in the roots to provide Al root exclusion. Even plants like eucalyptus trees can exude phenols that can bind Al and thereby prevent uptake [12].

Outside of chemical means there are also physiological means to confer Al resistance that plants can employ. This includes modifications to the root cell in order to change the intake or binding of Al^{3+} [13]. Al resistance mechanisms are largely dependent on exclusion of Al from being internalized, study of these mechanisms give us little information about the actual toxic effects of Al since Al never enters the tissue.

Identification of Aluminum mutants

It is of particular interest to understand how Al toxicity affects plants grown in these acidic soils, especially with regard to how Al actually halts root growth. Al^{3+} has the potential to bind to any anionic site, this includes but is not limited to: the cell wall, plasma membranes, and even DNA. It is also of interest to potentially use these new insights into Al toxicity to develop new strategies to provide plants with the means for tolerating the Al^{3+} to which they are exposed. This is an alternative strategy to confer plants with the capability to grow in Al toxic soils.

Al toxicity is thought to be extremely complex, especially when one considers that there are multiple anionic binding targets throughout the plant for Al^{3+} including sites such as the cell wall, lipids, proteins, and the phosphate backbone of DNA. In order to better understand how internalized Al affects plants, ethyl methanesulfonate (EMS) was used to generate a series of Arabidopsis mutants with increased Al sensitivity (*als* mutants) [14]. One mutant in particular presented a severe phenotypic response to Al^{3+} even at very low levels that had no effect on Col-0 wt roots. This mutant, *als3-1*, was determined to have a lesion in a gene encoding an ABC-like transporter (*ALS3*) that is homologous to bacteria protein YBBM and functional homolog of STAR1 in Oriza Sativa

(rice) is thought to be responsible for transportation of Al^{3+} throughout the plant's vasculature away from the growing root tip [14]. Loss-of-function *als3* mutants are thought to be unable to transport Al^{3+} to less sensitive tissues thus causing the extreme Al-dependent mutant phenotype, which includes terminal differentiation of the root tip, endoreduplication and enlarged cell size. Having all of these phenotypic changes occur at levels of Al that have little effect on Col-0 wt.

It is unlikely that these phenotypic changes that are being observed in the *als3-1* phenotype are driven by a single factor as they can be viewed as part of two normal processes under their normal functions. First, the cells would normally expand as part of replication and under normal circumstances they would divide. There are also many factors that are involved in this process such as hormone signaling (auxin polarization [15] for example), molecular signaling, and gene transcription that are all part of the process. Previously characterized as part of the *als3-1* phenotype is also the loss of the quiescent center (QC) which is considered to be stem cell niche. However, previous research showed that this loss has some connection to the activation of SOG1 by ATR, demonstrating that both *atr-4* and *sog1-7* in response to Al^{3+} does not lose its QC [16,17]. In addition, ETHYLENE RESPONSE FACTOR 115 (ERF115) is thought to be one of the main factors that regulates the division of the QC [18]. Therefore, combining with what has been seen with *sog1-7* would seem to suggest that SOG1 could be interacting with ERF115 to lead to the loss of the QC either directly or through other factors.

However it is not yet known what exact molecular factors are playing into this process in response to Al exposure. Though based on the presented evidence the current hypothesis points toward an ATR, and SOG1 driven pathway as a means of

attempting to maintain the overall genomic integrity of the organism. While the other likely possibility could be a form of hormone response which will be discussed later via either an auxin response. Of note is also that these two hypotheses are not mutually exclusive. SOG1 has many gene targets for which it is responsible for, and could be activating any number of suites of gene. To further examine this, RNA sequencing (RNAseq) is required to see specifically what genes are being regulated to SOG1 in response to Al exposure.

NRAT1 and ALS1 were studied in rice and Arabidopsis, respectively. These two proteins work together to remove Al from the cell and contain it in a root vacuole. This happens where NRAT1, which is a natural resistance-associated macrophage type protein, will export the Al out of the apoplast and into the cytoplasm, and ALS1 which is an ABC type transporter imports Al into a vacuole [9]. NRAT1 is a member of the NRAMP family of proteins, which transports ions across the cell wall. Unlike its other family members, NRAT1 specifically transports Al^{3+} but excludes other smaller divalent cations such as Mg2+ [19].

In an attempt to sequester the Al to areas of the plant that are less biologically important, Al gets transported from other areas such as the root to other parts of the plant such as the leaves for one example. Transporters functioning similar to ALS3, which is an ABC type transporter, are useful tools towards accomplishing this goal. Since Al can bind any anionic (negatively charged) site in the plant, those Al^{3+} ions that are not chelated by OAs or other Al resistance mechanisms in the root rhizosphere, will make their way to the cell wall of the root. It is here that the plant can change the morphology of the root cell wall, causing the Al to bind in an attempt to prevent the Al

from entering into the root system. This is done through the availability of pectins that whose structure has negatively changed carboxylase ring that binds the Al^{3+} and prevents it from entering the system [20]. For Arabidopsis specifically, it can also use hemicellulose which can bind Al the same as pectins. During cell expansion these hemicellulose are cleaved and rejoined, when the enzyme responsible for this process is knocked out, the plants were resistant to Al which suggests that these hemicellulose may play a role in Al tolerance for the plants. Additionally, proteins termed SENSITIVE TO ALUMINUM RHIZOTOXICITY 1 (STAR1) and STAR2, alter the composition of the cell wall, to inhibit binding of Al [15]. This is done by altering the amount of glucose in the cell wall which changes the availability of binding sites of Al and leads to less binding and more rigid cell walls.

Previous research has shown that SOG1 is a critical transcription factor that is activated in response to DNA damage and functions to actively stop root growth following Al exposure. It is important to note that this is not the only transcription factor involved when the plant experiences Al toxicity. The cis2-his2 zinc finger like *SENSITIVE TO PROTON RHIZOTOXICITY* 1 (*STOP1*) is another transcription factor that is key to Al response [21]. It was later characterized to activate other genes as well, including *STOP2*, *ALMT1* and *ALS3*. This gene seems very important as it has many orthologs among other species such as rice, wheat, and eucalyptus to name a few. The original characterization determined that STOP1 was important for both protons and Al toxicity [22] whereas other studies found that STOP1 functions more for defense in the acidic soils, but aids in the defence against Al toxicity by its regulation of STOP2 [23]. These two genes are very similar and there was difficulty determining which was more

responsible due to some overlapping and redundant functions. One of the most important genes that STOP1 regulates is ALMT1, which is upregulated both in acidic soils and following exposure to Al^{3+}.

Besides giving us insight into how Al is detoxified, the *als3-1* mutant has become a very useful tool for determining what factors play a role in Al-dependent stoppage of root growth and mechanisms of Al toxicity. This has been achieved through an *als3-1* suppressor screen, which was performed to find second site mutations that could reverse the extreme Al hypersensitivity of the *als3-1* mutant. For this screen, *als3-1* seeds were mutagenized with EMS and then grown in the presence of levels of Al that severely inhibit the mutant but not Col-0 wt. Suppressor mutants that could restore the growth of *als3-1* to levels comparable to wt or greater were subsequently isolated. In doing so, one mutant in particular was discovered that resulted in very high levels of Al tolerance even though the mutant had the *als3-1* mutation in the background.

This mutant was found by map based cloning to contain a single nucleotide change in *ATAXIA TELANGIECTASIA MUTATED AND RAD3 RELATED* (*ATR*) [17]. In brief, ATR is responsible for the regulation of the cell cycle in eukaryotes and detects DNA damage in the form of persistent single strand DNA (ssDNA) or replication fork stalls. Loss-of-function mutations in *ATR* suppressed the Al-hypersensitivity phenotype of *als3-1* and actually gave greater than wt root growth, thus suggesting that this mutation may be useful for increasing Al tolerance in plants in general. These findings provided the initial evidence to form the hypothesis that Al^{3+} was a DNA damage agent that activates an ATR-mediated cell cycle checkpoint response to halt root growth.

Additional mutants were identified as suppressor mutants of *als3-1* such as

Aluminum Tolerant 2 (ALT2). ALT2, when overexpressed lead to aluminum tolerance in the plants. This protein which is a WD40 protein responsible for ubiquitinating proteins, this process leads to a ubiquitin being added to a protein labeling it for degradation and therefore the labeled proteins function is suppressed. When created as a double mutant with *ALT2*;*als3-1* these plants grow larger than wild type and appear very tolerant of Al even when the plant is exposed to Al on a 1.5 mM gel soak environment.

Model of Molecular response

It is expected that ATR functions as part of a pathway to detect Al dependent damage and translated for stoppage of cell cycle progression to either promote DNA repair or differentiation of the root stem cells that form the quiescent center (QC). Previous work has shown that ATM and ATR responded to other DNA damage agents such as gamma (ɣ-) radiation or DNA crosslinking agents through a transcription factor that activates a suite of DNA damage response factors that function to repair DNA or if too damaged, force cells into a program of endoreduplication rather than cell division [24,25]. Recent work has shown that this transcription factor, *SUPPRESSOR OF GAMMA RESPONSE 1* (*SOG1*) is also required for stoppage of root growth following Al treatment since loss-of-function mutations in *SOG1* suppress the *als3-1* phenotype. In support of ATR and SOG1 working together to control response to Al, it was shown that ATR has the capability to phosphorylate SOG1 *in vitro* [26].

In support of Al^{3+} acting as a DNA damage agent, it was also found that Al treatment results in double-strand breaks (DSBs) in DNA as shown by Comet assays [16] as well as accumulation of micronuclei in *Vicia faba* root tips [27]. These findings are

particularly interesting when one considers these in relation to the predicted role of ATR in DNA damage response (DDR) and Al tolerance. While ATR is responsible for detecting ssDNA and replication fork stalls, it has not been linked to response to DSBs in Arabidopsis. To further complicate this, analysis of a knockout of Arabidopsis *ATAXIA TELANGIECTASIA MUTATED* (*atm*) shows that it does not suppress the *als3-1* Al hypersensitivity phenotype [26]. This conundrum leads to the questions of whether DSBs are a direct or secondary effect of Al treatment and what type of Al dependent effect on DNA activates the ATR-dependent pathway.

Using previously studied DNA damage response targets of SOG1 that are up-regulated in response to ɣ-radiation [24], it was shown by RT PCR that treatment with Al also increased expression of almost all of the same genes. Examples include *BRCA1* and *PARP2*, both of which encode DNA repair factors that have a substantial increase in expression following Al exposure. Justification of the argument that ATR and SOG1 function together to regulate Al-dependent changes in gene expression that promote terminal differentiation of the root, loss-of-function mutants for both *atr* and *sog1* result in failure to induce these DDR genes following Al treatment. This was not true for an *atm* loss-of-function mutant, further supporting the argument that Al toxicity is detected by ATR rather than ATM.

Combination of these observations leads to a current working model whereby Al enters the cell and affects DNA integrity or structure either directly or indirectly. Some possible direct effects could include binding to the phosphate backbone, creating DNA crosslinks or less likely but still possible is that Al^{3+} is binding to the actual base of the DNA and creating a lesion. Additional indirect effects can occur that include generation of reaction oxygen species (ROS) that can also lead to damage of the bases of the DNA which can lead to lesions. These interactions are speculated to promote formation of persistent ssDNA and/or cause the halt of progression of the replication fork, which causes the activation of ATR and SOG1 leading to cell cycle arrest. They can also lead to direct DNA damage, followed by the stoppage of cell cycle progression which will lead to an attempt to repair the DNA damage and continue the cell cycle. However if the damage cannot be repaired the cell transitions to endocycling whereby the cell skips the M phase of the cell cycle and instead enlarges and duplicates its DNA leading to terminal differentiation and severe increases in ploidy of those cells.

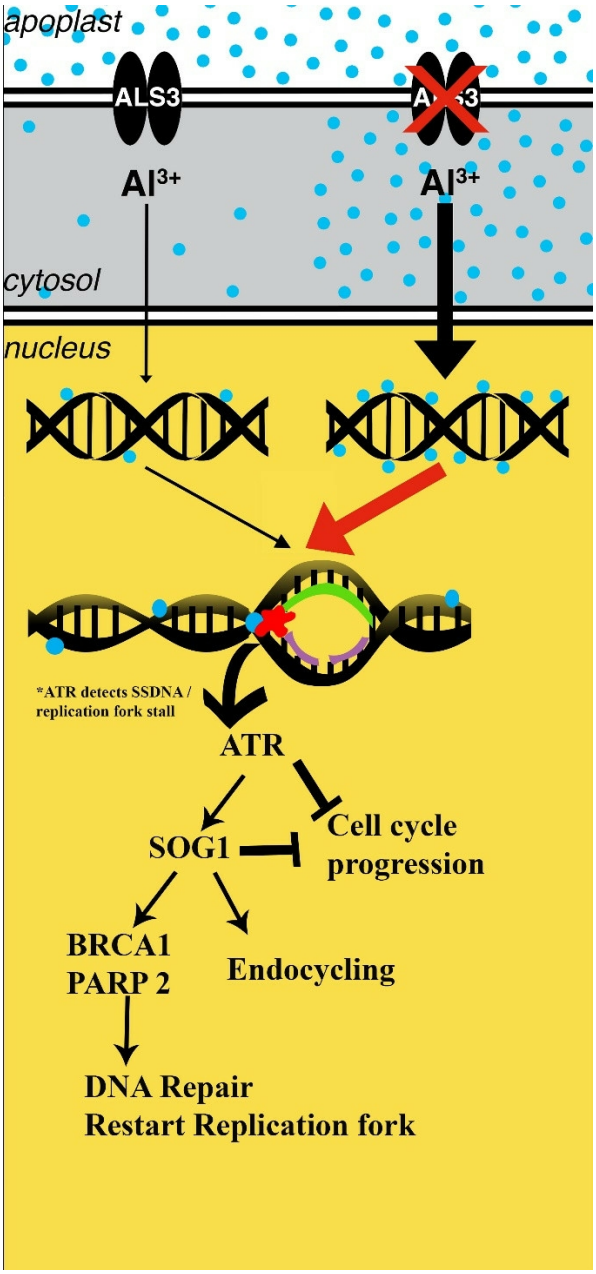


Figure 3: Current working model of the DDR to Al toxicity

Al^{3+} enters the plant vasculature, after making it to the nucleus, the Al^{3+} binds to the negatively charged back bone of the DNA, which leads to replication fork stalls / collapses and ssDNA and DSBs. This also triggers the recruitment and activation of ATR which halts the cell cycle and activates SOG1, which can also halt the cell cycle. SOG1 can either transcribe DNA repair factors such as BRCA1 and PARP2 to attempt DNA repair and restart of the replication fork, cause the cell to undergo endo cycling.

DNA Damage Response

Due to the complex nature of Al toxicity and the complexity of DNA biochemistry, determining the outcome of the DNA damage response has been quite difficult. Based on the literature, there are two possible links between Al toxicity and DNA damage response (DDR). Currently there is no evidence to support that Al^{3+} directly binds to DNA, however previous works has shown in response to Al^{3+} present in the plant a DDR is initiated. Due to this uncertainty about whether the DNA and the molecular machinery of the cell is affected by the presence of Al^{3+}, there are two likely situations that could explain DDR and detected DSBs. First, the Al^{3+} is not actually affecting the DNA or the molecular machinery and instead it is detected to start an unnecessary DDR that creates the DNA damage that is seen via the COMET assay.

The second possibility is that there is in fact some as yet unidentified damage occurring and the plant is rightfully trying to repair it in order to maintain genome integrity. Both of these hypotheses have merits, the first situation is supported by the evidence that when you KO the upstream regulators of the DDR, the plants do not undergo terminal differentiation or endoreduplication, and these KOs seems to grow as well or better on Al gel soak media than wild type. When ATR and SOG1 are both functional the plant undergoes DNA replication in the form of endoreduplication. Suggesting that these if there are replication fork stalls the plant has the means in which to either ignore or bypass them. This could be accomplished by firing other origins of replication, and various other means.

The second possibility however is not excluded from this same phenomenon, since these plants are also growing very small and sickly, showing there is likely significant damage to plant that it can barely cope with. There are many different types of mutants that have been used to characterize Al in previous studies, however the focus of this study will be on those factors that fall into two main categories of DNA damage: sensing and activation factors, and DNA damage repair factors. *sog1-7* for example is part of the former where ATR would sense possible damage and phosphorylate SOG1 to initiate a DNA damage response. When these type of factors are knocked out, the plant appears to grow normally and sometimes even better than wild type. In contrast, those mutants that are involved in the repair of Al-dependent DNA damage when knocked out lead to a plant that is very sensitive to Al. Loss-of-function mutations in factors such as BRCA1, LIGIV, PARP2, and other repair enzymes are examples of the latter.

To further understand the pathway of the DNA damage response shortly after the activation of ATR and SOG1, various reliable tools have been used to determine the occurrence of DNA damage. Using molecular tools, such as the comet assay and phosphorylated gamma H2AX assay researchers can perform experiments . Early studies done by Rounds and Larsen [17] showed via the COMET assay that there was damage occurring based on the increase of the micronuclei that are observed to make up the “comet tail.” However, due to the nature of the comet assay, the knowledge and way it was performed at the time can either be single stranded DNA or double stranded breaks of the assay under neutral conditions as part of the micro nuclei shown in the tail after the loss of the nuclear matrix [17,28]. The latter does not have to have both , it could

just be double stranded breaks since it is merely dependent on pH at which the assay is run. These outcomes can arise due to numerous reasons: a replication fork stall can lead to its collapse, which will create a double stranded break and possibly lead to an increase in the amount ssDNA will then be persistent. While the results of the COMET assay are inconclusive in terms of what kind of damage is occurring, it does demonstrate that with treatment with Al^{3+} there is an increase in micronuclei meaning an increase in DNA damage. From the research done by Sjogren et al. [26] PARP1 and PARP2 showed sensitivity to Al. Sensitivity progressively increased as double mutants (*parp1;parp2)* and even more so for the triple mutant which included *ku80, parp1;parp2;ku80*. PARP1 and PARP2 played a major role in Microhomology Mediated End Joining (MMEJ), which is referred to as alternative Non-Homologous End Joining (alt-NHEJ) [29], whereas, KU80 is a primary factor of canonical NHEJ. Based on the growth data it suggests that loss of these processes in the presence of Al leads to plants' inability to correct molecular issues that occurred as a result of the Al, thus resulting in Al hypersensitivity.

DNA Damage Repair Pathways

The primary repair mechanism for repairing DNA damage that is used by Arabidopsis is Homologous Recombination (HR). HR uses a sister chromatid to make near perfect repairs using an undamaged copy of the DNA. This occurs by identification of the homologous region on the sister chromatid, followed by strand invasion of the damaged area creating the structure known as a Holliday junction [30]. The damaged area is then removed and repaired with the sister chromatid serving as the template. Finally, once the repair is complete, the Holliday junction is resolved and the two sister

chromatids return back to their normal state [30]. Unfortunately, there is a very limited window during the cell cycle in which HR can occur since this process is dependent on the presence of sister chromatids, which are around only during the actual process of DNA replication.

Repair processes such as NHEJ and MMEJ differ from HR in that instead of repairing the DNA using a duplicate undamaged region of the DNA on a sister chromatid as a template, the molecular machinery for these processes functions independent of a template. In these processes, the machinery creates a cut depending on the type of of the repair that is necessary with the possibilities including a DSB (NHEJ, MMEJ) or single strand knicks (Base Excision Repair (BER) or Nucleotide Excision repair (NER)). Each of these processes lead to different outcomes.

NHEJ and MMEJ are more prone to creating insertions and deletions (INDELs) which can lead to issues in genomic stability. NHEJ simply leads to excision of the damaged area and ligation of the two blunt ends of the DNA to repair the damage without concern for fidelity. MMEJ on the other hand uses microhomology driven repair when there is homologous sequence of 5-25 base pairs (bp) within the same strand of DNA on either side of the damaged area of DNA. This repair leads to a double stranded break, but in the process of pairing the microhomologies more information can be lost due to deletion of intervening sequences. However, as Schuerman et al. [31] points out these errors in DNA repair can lead to essential genomic variability within a plant population, possibly contributing to evolutionary change.

Even though BER and NER can perform near perfect repairs since they simply remove one base and replace it with the corresponding nucleotide they have their

drawbacks as well. While BER is responsible for repairing only the affected base, it is characterized as having both long and short repair mechanisms that can affect anywhere from just one base up to 10. During this process, single stranded breaks are generated, which can lead to the affected area being targeted by other repair mechanisms [32]. NER occurs in two different forms- either throughout the genome or coupled with transcription called transcription coupled repair (TCR).

This process takes place when the RNA polymerase encounters a legion on the DNA strands it is trying to transcribe, which leads to a stall in transcription. Various factors are recruited including RPA, XPA, and CSB, together with RNA polymerase will lead to an incision in the DNA and removal of the legion, followed by repair and continuation of transcription [33]. Global NER on the other hand detects DNA lesions via XPC-RAD23B which can be bound by TFIIH. Thus, this leads to the use of the associated subunits to trace the DNA and detect the lesion via stalling [34].

After the lesion is verified via stalling this allows the recruitment of XPA, XPG, and RPA. ERCC1-XPF interacts with XPA to create an incision that leads to resectioning of the DNA via polymerase δ/ε/κ, with LIGIIIα/XRCC1 (LIGI can also be used) to seal the remaining knick [34]. In both cases it occurs when lesions, which are sites of damage to the nominal structure of DNA or to the base pairing of the DNA is detected on the DNA. While this definition is very vague it can apply to many types of damage including but not limited to mismatch bases, DSBs, and intrastrand crosslinks. In both cases NER normally leads to the removal of approximately 30 bp during the repair which are later synthesized by the polymerase that functions after the excision .

Mismatch repair (MMR) as the name suggests is responsible for the identification and repair of mismatched nucleotides. However, it is also responsible for ensuring proper homologous repair and preventing homeologous recombination from taking place, a process by which MMR prevents recombination from taking place with a reference sequence that is similar but not identical [35]. In terms of the types of damage it repairs, MMR has been classified to respond to interstrand crosslinks, UV photo products, alkylation, and oxidative DNA damage [36]. Even though the system of MMR is highly conserved, it has many parts and many repair mechanisms have overlapping function. Studies have been done for MMR in mammals, yeast, and bacteria, but there is still little known about which differences exist within the realm of plants. There have been some studies [37–39] to measure the mutation rate for mutant plants that are deficient in MMR, to which those researchers found a selectable phenotype [37].

Repair of alkylated bases via DNA Alkyltransferase uses a protein called Alkylguanine DNA alkyltransferase (AGT), which is responsible for identifying bases that have been alkylated and removing the alkylation using its cysteine residue. This process has been most noted in bacteria and mammals, but currently has not been determined for plants. However, a process such as this one could be what repairs the DNA if possible while transcription is taking place to remove any alkylated bases due to Al toxicity via the generation of excess ROS which could lead to oxidative stress and alkylation of DNA bases. The base itself can be oxidized and lead to apurinic/apyrimidinic site (AP site) [40] as well as the oxidative damage can lead to alkylation [41].

Since we know already that one of the possibilities of the DSBs that were observed could be due to the collapse of a replication fork, this could explain how the replication process could cope with Al. AGT can be used up in this process as it is irreversible and leads to the protein being degraded [36]. It has been hypothesized that BER could substitute for the lack of AGT in the plants. Due to the complexity of the DDR, identifying a specific repair pathway that is responsible for the DDR with DNA repair following exposure to Al^{3+} has been extremely difficult.

This is also the basis for the reporter line mentioned previously, IC9, which has a broken beta-glucuronidase (GUS) gene [42]. This gene is fragmented in two parts, one on each sister chromatid, which allows scientists who wish test the prevalence of HR from a treatment to perform an experiment with the treatment of choice, in this case Al, followed by fixing and staining the plant tissue to see blue spots. These blue spots are created by HR taking place to repair the broken GUS gene where by having the gene expressed, and will show as blue spots on the plant tissue. However the shortcomings of this approach are that the blue spots of quite rare, and in order to get a proper population size to make judgements about the state of HR by the treatment requires hundreds of the plants and screening looking for blue spots. While this technique presents promise for helping to provide evidence for the involvement of HR, there has yet to be any published use of this technique with in the study of Al toxicity.

Rogakou et al [43] showed in Arabidopsis that gamma phosphorylation of the histone, H2AX, is correlated with the formation of DSBs not just in humans but also in plants thanks to a conserved SQ motif at serine 139 of the histone. This has allowed the phosphorylation of H2AX to become a useful tool to monitor for the presence of double

strand breaks. In the case of theorized models for Al toxicity it would fit that ATR and ATM can both phosphorylate H2AX [44]. While H2AX phosphorylation is a good indicator of double strand breaks it does not indicate much else in the way of how the organism is handling the response. Future studies could use this ɣ-H2AX phosphorylation to determine if and quantify the level of DSB formation in response to Al^{3+} exposure.

Response to Al toxicity has other factors involved beyond just the Al induced DDR. The plant has a whole host of genes whose expression is changed in a more general way such as the pathogen defense response and (ROS) as well as hormones that change how the plant grows when it detects these kind of problems in the biological system [45].

With regard to hormonal control of response to Al, both auxin and brassinosteroids have been found to have essential roles. While these hormones have general roles in plant growth and development, they take on specialized responsibilities in response to Al. Panda et al. [15] discuss that Al directly affects auxin transport, which is responsible for control of cell elongation, and brassinosteroid response, which are responsible for cell division and expansion. The other role that hormones play can have to do with gene responses in terms of their regulation which could have to do with gene regulation in response to Al that would independent of SOG1 yet still responding to the toxic stress of Al.

It is interesting to note that a plant's response to Al is very similar to what happens to when a plant is grown in a low pH environment. In this environment, excess protons can not only lead to DNA alkylation and activation of the DDR as noted earlier but also leads to the plant mounting a ROS response [40]. ROS buildup would be the

result of the plant generating ROS to bind to free protons or atoms with positive charges, such as Al^{3+}, to negate or minimize the destructive effect that these positive charges can have. This is important since if Al^{3+} is left in its cationic state, it would not only bind to DNA as previously stated but also could displace Mg2+, which plays a role in stabilizing and maintaining DNA structure through electrostatic interactions.

If this occurred, it would be predicted that the non-covalent interactions would cause conformational changes that could lead to the constriction of DNA and slowing or halting of the replication fork. This would lead not only to a damage response, but also cause issues with transcription and overall access to the genetic informations of the DNA. This sort of behavior has been shown to happen with heavier atoms such as cobalt, which in its toxic form also have a trivalent charge [46].

Analysis of the Transcriptional Response to Aluminum in Arabidopsis

Introduction

Following exposure to severe aluminum toxicity, plants will undergo morphological changes that are associated with stoppage of root growth in a SOG1 dependent manner [26]. As SOG1 is a key player in the transcriptional response to DNA damage, this suggests that these morphological changes are at least dependent on altered expression of genes that control this phenomenon. Previous research done on the *Arabidopsis* transcriptional response to aluminum found aluminum inducible genes which are dependent on functional SOG1 [26]. The prior approach used a reference gene that was found in this experiment to be differentially expressed with the applied treatment of Al^{3+}, this lead to a biased survey using relative expression that attempted to tie previously described SOG1 transcriptional targets to Al dependent changes in gene expression. With advances in Next Generation Sequencing (NGS), the capture and sequencing of mRNA can be accomplished to survey in an unbiased manner which SOG1 regulated genes are expressed in an Al-dependent manner.

In the previous work by Sjogren et al. the aluminum sensitive mutant, *als3-1*, and its suppressor mutant *sog1-7* were used for phenotypic studies and qPCR (quantitative real time PCR) analyses to test for a relationship between Al toxicity and DNA damage. In this study a RNASeq approach was used to provide the basis of a less biased and exploratory study to determine potential key factors for SOG1-dependent terminal

differentiation of the root tip following Al treatment. This is accomplished by using the whole transcriptome for normalization vs one gene as reference. These loss of function mutants along with wild type Arabidopsis provide the means by which a targeted transcriptional comparison can be performed. Using a hypersensitive mutant, along with a tolerant mutant the results can be focused on target genes that are responsible for stoppage of root growth in an Al^{3+} dependent manner.

These experimental variables were chosen to correlate the results of the RNAseq experiment with those of the previous RT-PCR experiments done by Sjogren et al [26] with the goal to determine differentially expressed genes that could be linked to the phenotypic changes described in previous research, with a secondary aim to verify the results of the previous study in a more unbiased manner. However there are major differences between this RNAseq experiment and previous research done with the transcriptional response with Al^{3+} and SOG1. Previously the research using qPCR took total RNA and used poly dT to create cDNA (complementary DNA) for expression analysis. Other RNAs such as small non coding RNAs or rRNA were not removed, though could be negligible after completion of the cDNA synthesis. Additionally, the RNAseq experiment was done in a stranded manner, leading to high sensitivity and less off target noise from expression of any antisense genes.

With these differences in mind, moving forward with the experiment, the previous work is treated as a proof of concept that there are genes that can be identified as target genes that are differentially regulated and even suppressed using *sog1-7*. Using this information this study is treated as an unbiased survey of the transcriptome to identify potentially new targets in another manner to elucidate the process of SOG1 dependent termination of root growth post exposure of Al^{3+}

Experimental Design

In setting up the experiment using a wild type along with both an aluminum tolerant mutant and a hyper-sensitive mutant, trends can be identified to correlated the mutant phenotypes of *als3-1*, which has severe Al hypersensitivity phenotype. *sog1-7* on the other hand, which is Al tolerant and should not display increased expression of these genes if indeed they are SOG1-regulated in an Al-dependent manner. If there is indeed genes that are Al dependent and required for the inhibition of root growth, then by identifying those genes that are upregulated in wild type and are also severely up regulated in *als3-1*, but are not up regulated or down regulated in *sog1-7*, then these genes could provide the targets for future research in to the mode of action as to how Al^{3+} causes the stoppage of root growth phenotype in a SOG1 dependent manner for Arabidopsis.

In order to test the expression of the various plant lines in a way where the expression would be as uniform as possible, it was necessary to grow the plants on gel soak media in a way that all three lines would be grown the same plate. To test the effect of the Al and maximize the difference in expression the control dose of 0.0 mM

Al^{3+} and and the treatment dose of 1.5 mM Al^{3+} were used. One additional concern mentioned above was to stay consistent with the previous research done by Sjogren et al, to do this the plants were grown in a growth chamber at 25 C for 3.5 days before harvesting the tissue (whole seedlings) and flash freezing it in LN_2 for later RNA extraction. Each experiment contained 3 technical replicates, with 3 total biological replicates performed to generate the results mentioned in the following section. Each plate contained approximately 100 seedlings of each line (constituting 300 seedlings per plate). The seedlings of each genotype for each plate were pooled together for each sample that would later have its RNA extracted.

After the RNA extraction the samples would undergo polyA capture by using oligo dT beads, with the resulting mRNA being prepared in to a stranded RNA seq library, and run on an illumina HiSeq to generate sequencing data for transcriptomic analysis.

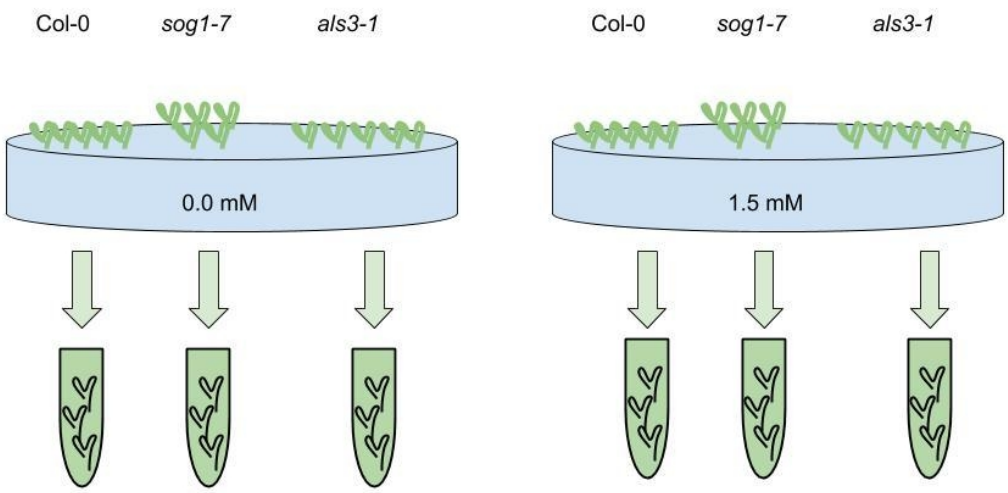


Figure 4: RNA Seq Experimental Conditions and Tissue Generation
To maximize the consistency of each line for each replicate both technical and biological, the three Arabidopsis lines were grown together the same gel soak media plates. All three lines would then undergo the same growth conditions either 0.0 mM or 1.5 mM Al^{3+} for 3.5 days, before having the tissue (full seedlings harvested and pooled by genotype) for downstream total RNA extraction.

Results

Using the experimental set up, it was very clear that differences between the treated and untreated existed. This would come as little surprise since applying the Al^{3+} treatment especially to *als3-1* shows a very distinct phenotype. In order to determine a set of genes that satisfy the requirements set forth in the hypothesis to be candidate genes involved if not responsible for the stoppage of root growth in a SOG1 dependent

manner in arabidopsis. Before being able to compare between samples, the sequencing reads of each sample had to be mapped to the transcriptome, counted, and normalized.

This was done using the TAIR10 reference genome and the corresponding TAIR10 annotation file (gff file) which in combination provides sequence regions of the transcriptome. Reads mapping to exonic gene regions are then counted, in this case only if they are on the correct strand, and then normalized within each sample. Then each counted and normalized sample is compared between treated and untreated samples of the same genotype. Since there are only two treatments in this study the labels applied by edgeR are from the target file provided, this deontes the control samples grown on the 0.0mM media as “0” and those that are treated are labeled as “ A” for the applied Al treatment on the 1.5 mM media the venn diagram provides an example of this (Figure 4).

After comparing between the control and treatment for each genotype differentially expressed genes (DEGs) can be determined using statistical tests based on the negative binomial distribution. In order to achieve a higher level of confidence in the results a stranded RNAseq kit was used, because of this when the analysis was done the counting was also done in a stranded manner, before the initial normalization by reads per kb per million mapped reads (RPKM). This provides a rudimentary normalization, the discovery of DEGs was performed with the GML method of the edgeR package which has its own normalization, seperate from the naive RPKM done separately in R. [47] .

For identifying DEGs, a fold change cutoff of 2 and a false discovery rate (FDR) of 1% were used. This resulted in the identification of 172 DEGs after exposure to Al by

comparing treated vs untreated samples for wild type and each mutant genotype. These 172 genes represent the union of genes that were differentially expressed (DEGs) in at least one of the genotypes, when comparing the untreated gene expression to the gene expression of the same genotype when the treatment is applied, the Venn diagram (Figure 4) shows the breakdown of these 172 genes as they compare between the genotypes. Of these 172 genes, 136 DEGs were found in *als3-1*, 40 were found in *sog1-7*, and 45 in Col-0 wild type.

Of the 172 DEGs, 9 genes had the same regulation pattern, where the same DEGS will go up or down with *als3-1* and Col-0 but not for *sog1-7*. Meaning that these same genes are up (red) or down (blue) regulated for both genotypes in a SOG1-dependent manner as shown in the venn diagram that follows (Figure 4). The figure also shows that 108 of these DEGs are upregulated when compared between treated to untreated samples leaving 64 DEGs that are downregulated or suppressed as a result of Al^{3+} exposure. In considering the three genotypes the hypersensitive mutant has the highest number of DEGs though, at the same time the majority of these genes are exclusive to als3-1 and while they provide potentially valuable information about the asl3-1 genotype, they do not contribute any information pertinent to the question at hand. sog1-7 and Col-0 were relatively the same in terms of order of magnitude of DEG counts. sog1-7 being al tolerant, and having the SOG1 transcription factor had the lowest total of DEGs which logically makes sense for that genotype.

In order to broaden the search results, manual curation of the list of DEGs was performed based on the following pattern of change: differentially expressed in wild type, more extreme change in *als3-1* and a suppression of the change in *sog1-7*. A group of

14 Al-inducible genes that were not SOG1 regulated were found in this RNAseq analysis, in which their expression was uniformly changed based on the exposure to aluminum regardless of the genotype. While these targets are no doubt important to Al response, a goal of the RNAseq experiment was to find those SOG1-regulated genes that are responsible for causing terminal differentiation and endoreduplication following Al treatment. However, they present an interesting targets for future research in to other facets of Al toxicity that could help provide answer to other facets of research on the topic.

The overall expression patterns of all the samples are presented in a heat map format in Figure 5. As a note, the Venn diagram (Figure 4) only shows those genes that were lower than the 1% threshold for the FDR with the results broken up by genotype and the subsequent overlaps. In contrast, the heat map gives an overall picture of the findings, since it shows all 172 genes presented in the Venn diagram and bar plots. Using the heatmap combined with Venn diagram results many genes that were found to meet the criteria of differential gene expression did appear to overlap between genotypes. The possibility exists that relaxing the FDR threshold to 5% could increase these overlaps. However, in doing so would also lead to larger total pool of genes with an increase in false positives.

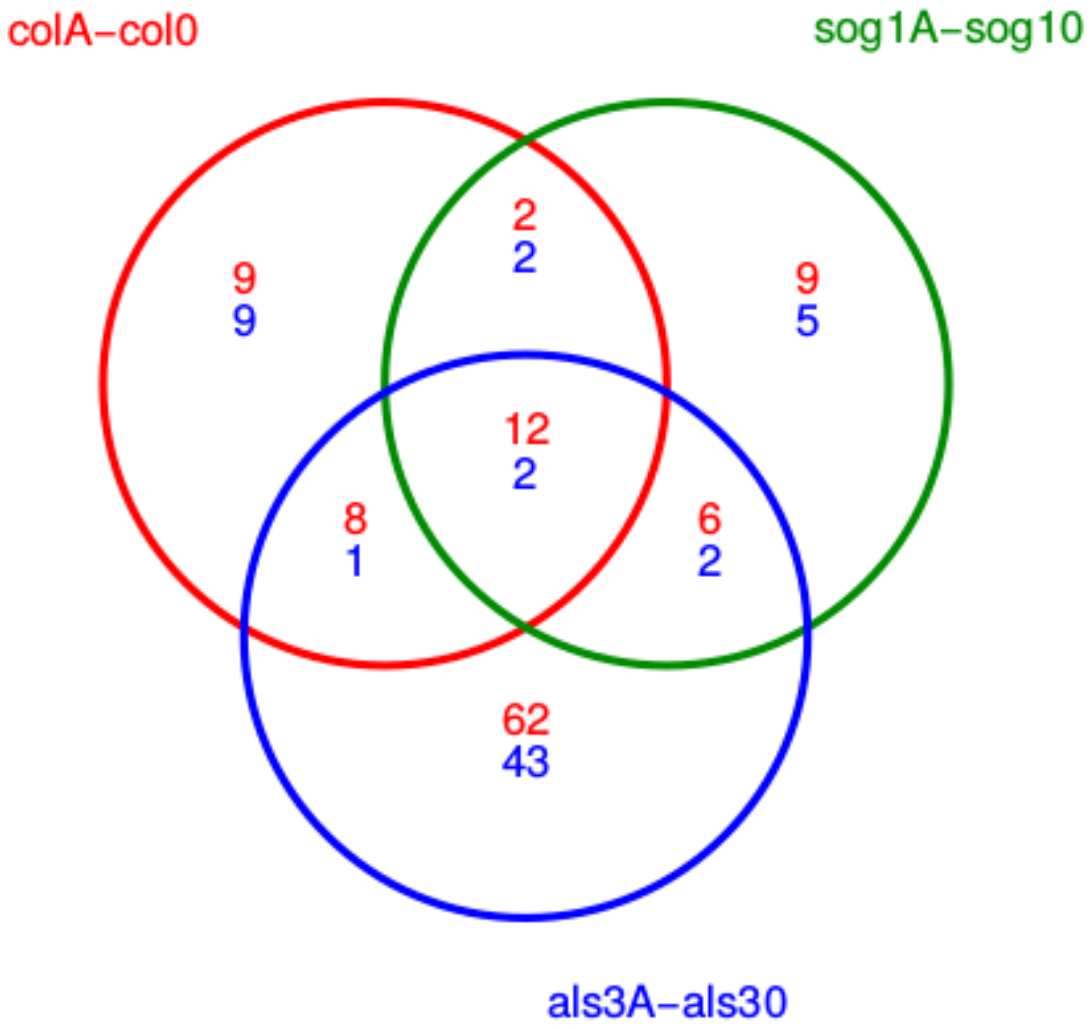
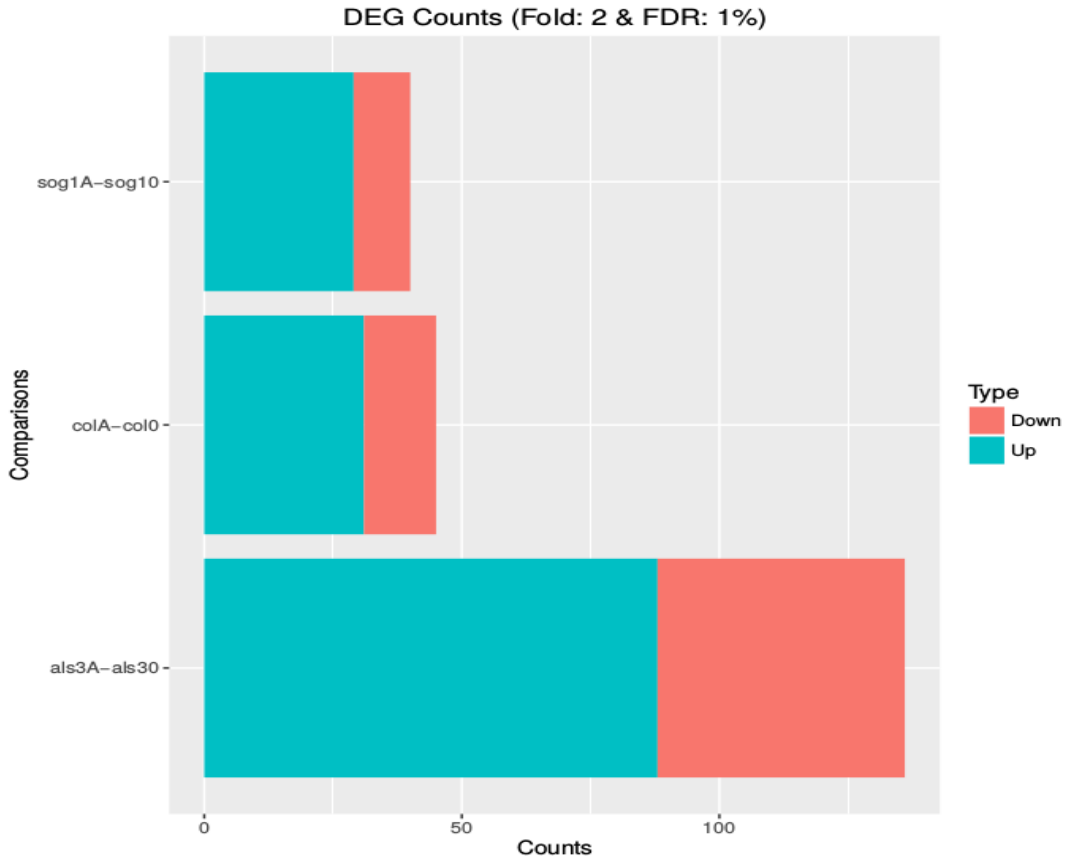


Figure 5: Venn Diagram of DEG Results
Venn diagram showing the overlap among DEGs sets meeting the confidence cutoffs of ≥ 2 fold change and an FDR of ≤ 1% FDR. The overlapping portions of each circle show genes that are in common between the genotypes of Col-0 (col), *sog1-7* (sog1), and *als3-1* (als3). Each circle or color is denoted by the comparison done to identify those DEGs within genotypes. This shown by the genotype "A" denoting treatment with 1.5 mM aluminum and the genotype followed by "0" to indicate samples grown on the 0.0 mM gel soak media. This indicated the comparison was done by the treated samples - the control samples to look at the differences to determine the DEGs.



Comparisons	Counts_Up_or_Down	Counts_Up	Counts_Down
colA-col0	45	31	14
sog1A-sog10	40	29	11
als3A-als30	136	88	48

Figure 6: Count Comparisons
Here the comparisons are listed by the DEGs comparisons denoted by the genotype followed by “A” representing treatment with 1.5 mM aluminum and the genotype followed “0” to indicate samples grown on the 0.0 mM gel soak media for each genotype genotypes of Col-0 (col), *sog1-7* (sog1), and *als3-1* (als3). A.) This figure shows a stacked bar graph presenting the counts of upregulated genes (blue) and the down regulated genes (pink) stacked to show the total Differentially Expressed Gene (DEG) count for each genotype. DEGs are defined as those that have a 2 fold or greater induction following Al exposure and a FDR of 1% or lower. B.) The actual numerical counts for DEGs following Al treatment of Col-0 wt, *als3-1* and *sog1-7*.

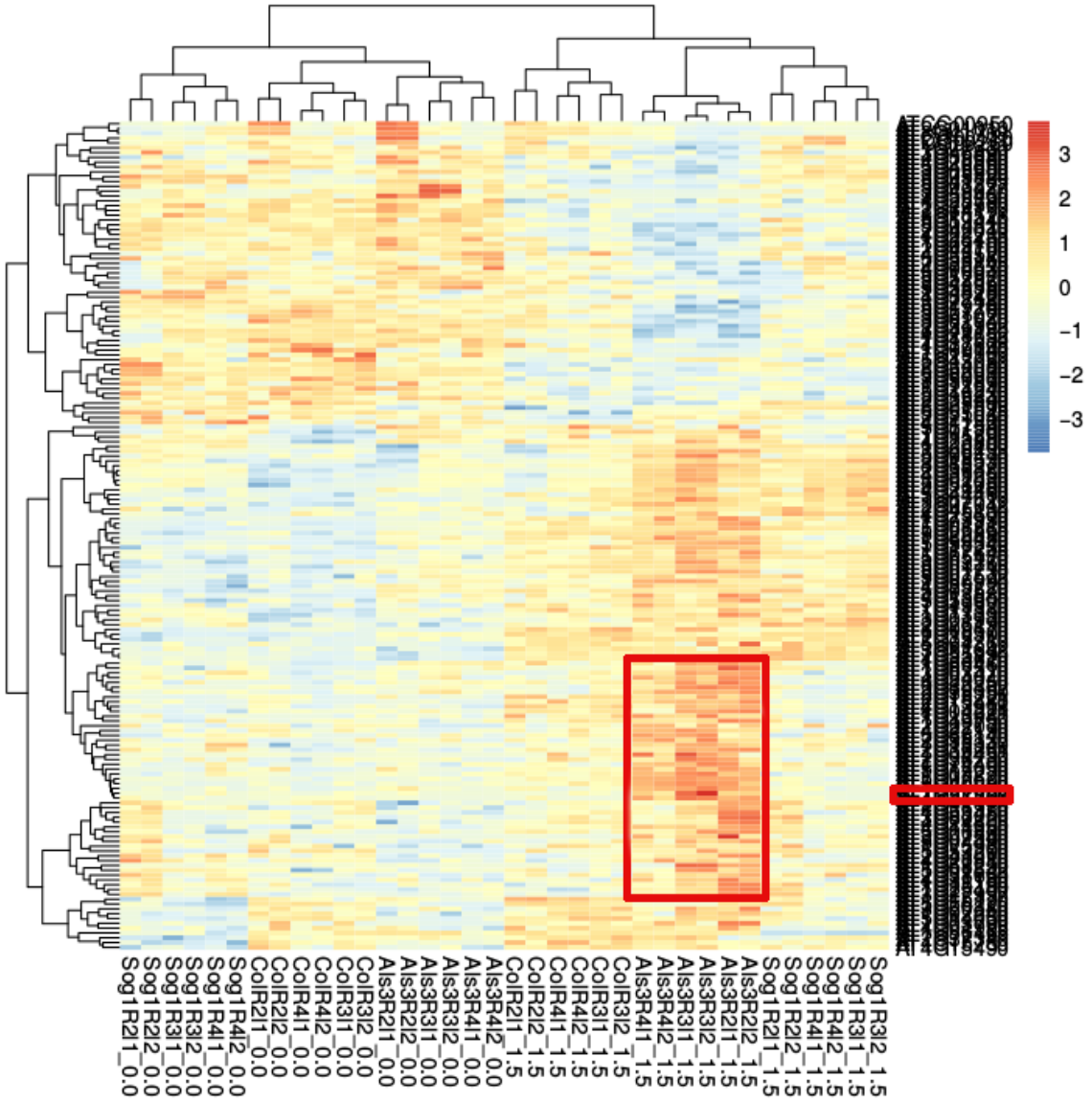


Figure 7: Heatmap of DEGs Correlated by Gene Expression
The heatmap generated from the union of all DEGs that met the criteria of ≥ 2 fold change and ≤ 1% FDR, with the genes of interest marked in large red square as the markers from *asl3-1* that are highly upregulated while expression is unchanged in *sog1-7*. The smaller red box denotes the location of AT5G07620 in the list of gene numbers, this gene was found to be strongly upregulated for *als3-1* in all the replicates, and followed the expression pattern expected of a gene that is SOG1 reulgated. The left hand side shows the correlation of each gene based on expression, and the top dendrogram shows the correlation between samples. Each sample is listed by its genotype of Col-0 (Col), *sog1-7* (Sog1), and *als3-1* (Als3) followed by it the replicate number R#, and finally by which lane of the sequencer the sample was run on.

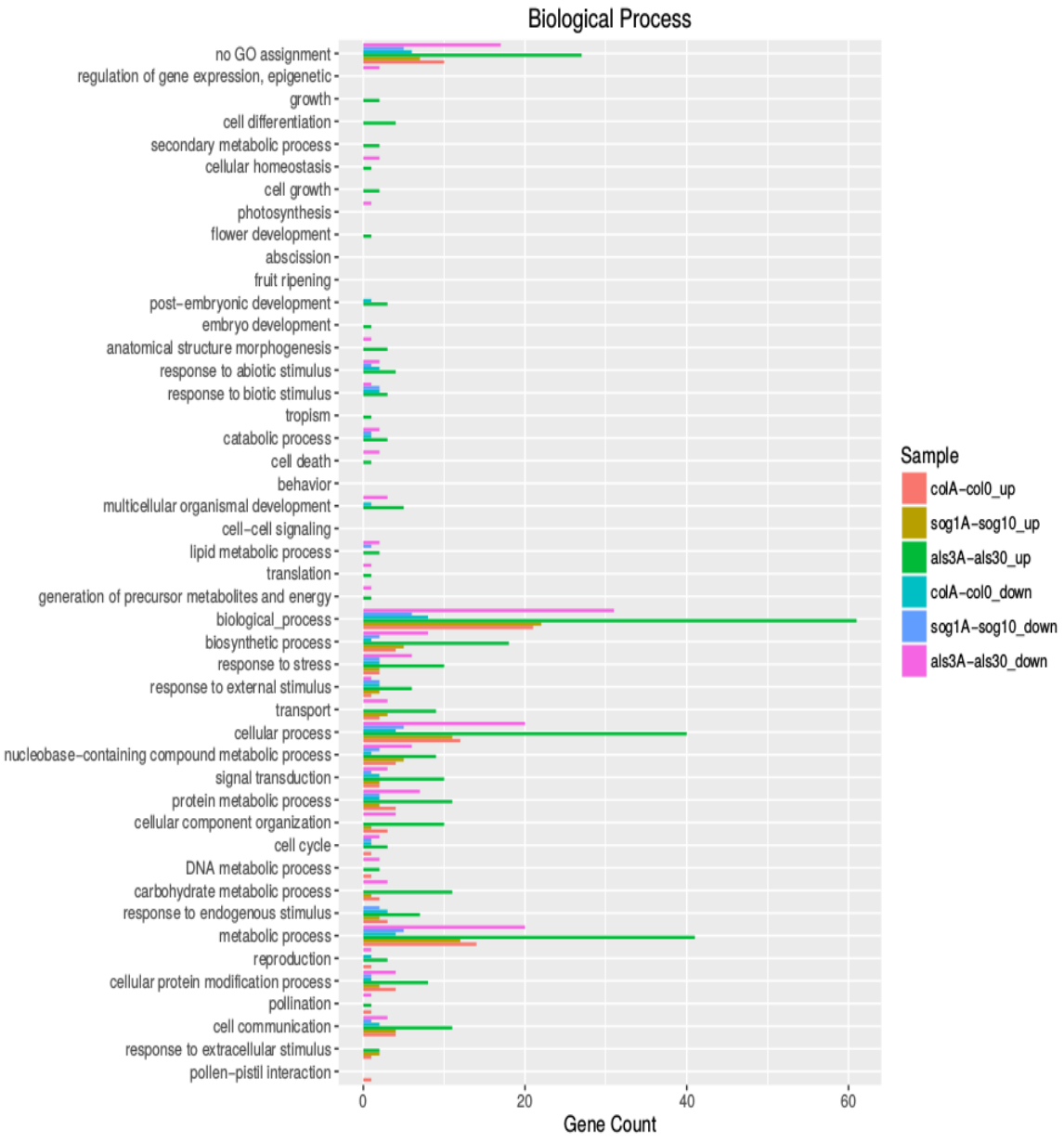


Figure 8: GO analysis using ensemble database - Biological process
Analysis of the GO information for the biological process demonstrating that many of the DEGs fall into very general categories. This is due to using the GO slim database, no terms were determined to be over represented using tools like GOrilla [46]. However further refinement of this GO analysis could be done in the future to find results that were missed. This is list is based on all DEGs not just those that were determined to be Al induced and SOG1 dependent.

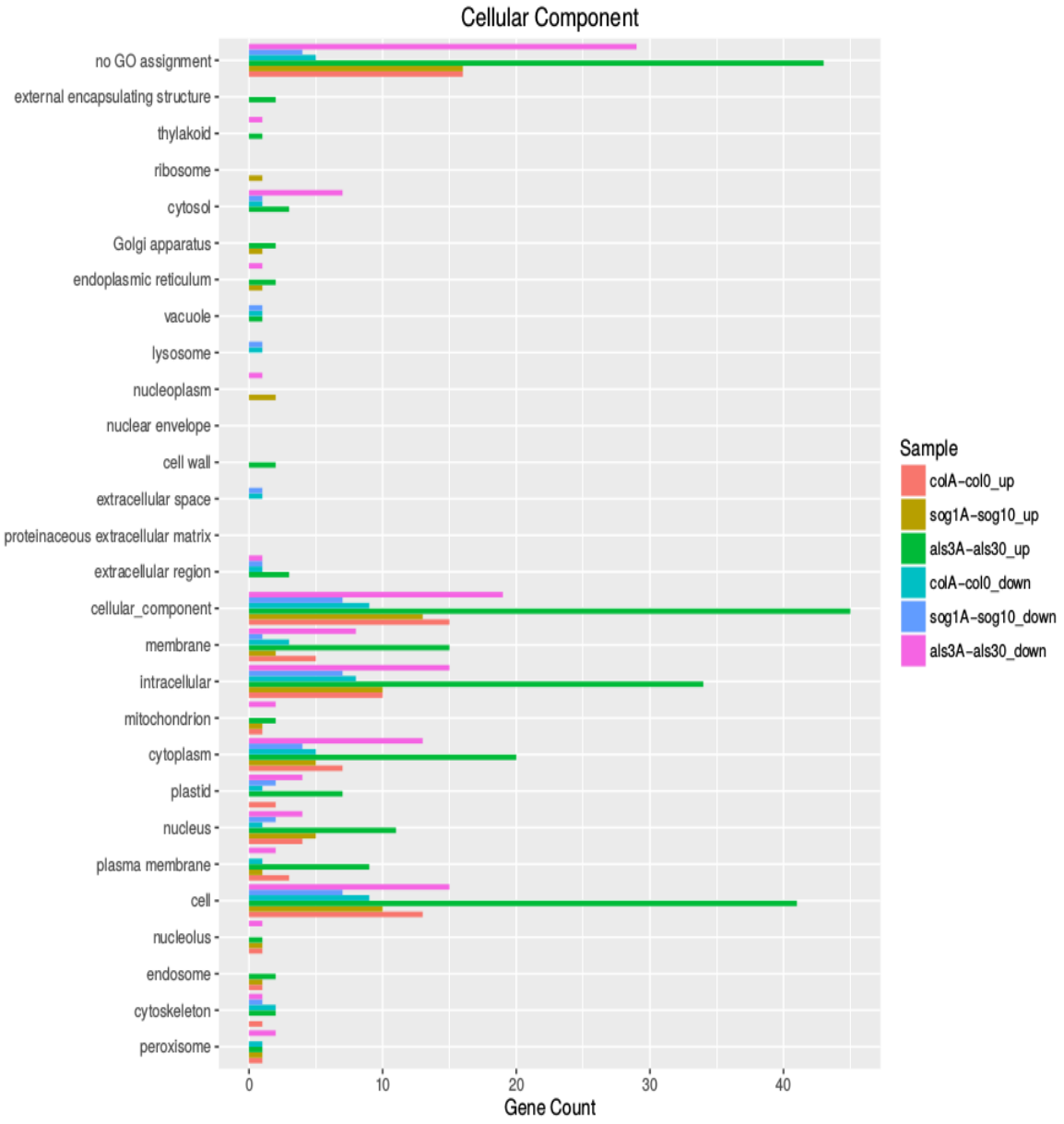


Figure 9: GO analysis using ensemble database - Cellular Component
Analysis of the GO information for the cellular component demonstrating that many of the DEGs fall in to very general categories.This is list is based on all DEGs not just those that were determined to be Al induced and SOG1 dependent. The cytoplasm and and plasma membrane, membrane, and nucleus, appear to be the key areas of the cell with Al treatment. Using tools online tools there was no GO enrichment determined, however future projects might be able to use the data generated to find additional details which could improve the results.

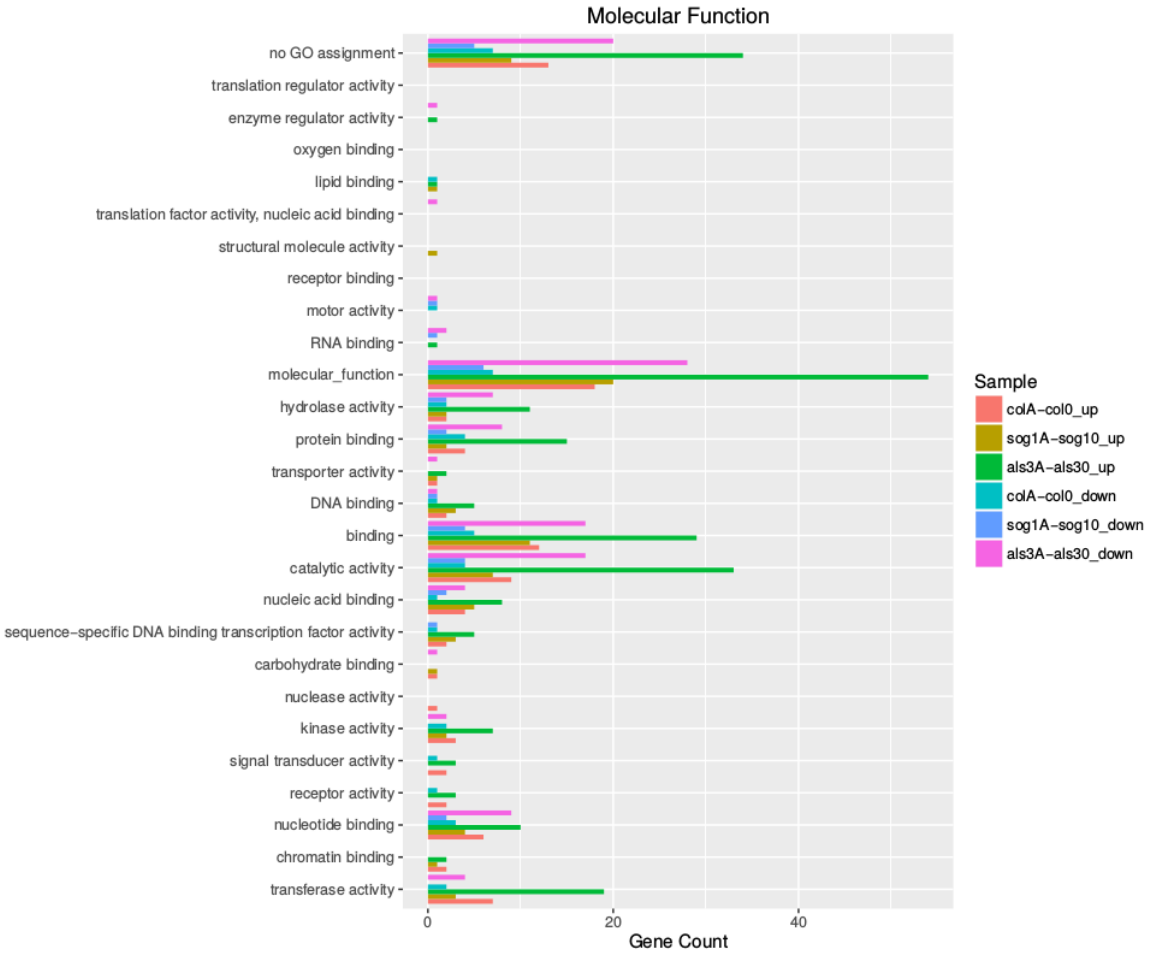


Figure 10: GO analysis using ensemble database - Molecular Function
Analysis of the GO information for the Molecular Function demonstrating that many of the DEGs fall in to very general categories. This list is based on all DEGs not just those that were determined to be Al induced and SOG1 dependent. Outside of the general categories there seems to be many genes that fall into categories: catalytic activity, binding and transferase activity. Using tools online tools there was no GO enrichment determined, however future projects might be able to use the data generated to find additional details which could improve the results.

To explore whether functional trends were present in the set of DEGs, Gene Ontology (GO) analysis was performed to identify the categories for DEGs to improve the understanding of the molecular response what type of genes played a role in the plants response to exposure to Al^{3+}. A secondary hypothesis is that plants mount a DNA damage response predicting an enrichment of GO terms related to this response. Since these analyses are dependent on GO annotations of Arabidopsis genes and remains somewhat limited, many of the differentially expressed genes are labeled simply as "no GO assignment" or with a generic heading annotation. Overall, the general trends of the GO terms show evidence that the majority of the genes are annotated as having substrate binding, catalytic activity, or transferase activity. With regard to cellular location, the GO assignment for location shows most are predicted to be cellular components that are intracellular, likely found in the cytoplasm.

Based on these drastic phenotypic changes noted in previous research [26] it is probable that many of genes being induced are part of signaling and cascade response to Al^{3+}. This would also include *At5g07620* which based on its predicted protein structure would have likely play a role in signal transduction, but does not possess a receptor domain. Overall it would make sense logically that to lead to a hypersensitive phenotype like *als3-1* that there would likely need to be a large change in signaling. One goal of this study was to identify genes that were directly responsible for the phenotypic changes observed in response to Al^{3+} exposure. Instead new targets were identified to help investigate and understand the pathway by which the plant responds to this stress. Due to the currently limited utility of GO assignments, categorization of many of the genes in relation to "Biological Process" and "Molecular Function" has not been informative. For

those DEGs that could be categorized, many fall into other general categories, such as metabolic process. Upon performing GO term enrichment, there was no significant results found.

Manual curation was performed on the list of DEGs in which the top 100 targets that showed significant fold change in *als3-1* were then filtered to identify those that demonstrated the transcriptional pattern described previously to be determined as SOG1-dependent changes, with a few that were independent based on the RNAseq analysis. The regulated pattern of expression was rank ordered based on fold change expression for *als3-1* (highest), Col-0 (intermediate), and then *sog1-7* (no change).

While there are genes that display an opposite pattern, those that are highly upregulated in the sensitive mutant *als3-1* but not in *sog1-7* were the focus of this study. Attributions of genes identified in this manner were determined by comparing gene ID's to listings in the database contained at The Arabidopsis Information Resource at www.arabidopsis.org (TAIR) [49]. The following information was retrieved from the database for each gene: GO annotations, a gene name if one exists, as well as a description of its documented or predicted functional role in the plant. Additionally, predicted/demonstrated tissue localization of the identified factors was used for determining which factors might be relevant to the stoppage of root growth following Al exposure. From this list of potential candidates that seemed to have roles in response to Al exposure, several were selected for further validation and testing via qPCR.

qPCR was performed to verify the results of multiple targets from the RNAseq survey (Figures 12, 14-20), with most of these genes chosen based on fitting the desired expression pattern outlined previously (Table 1). Those genes that are upregulated in

als3-1 but not *sog1-7* could be genes that are part of a SOG1-dependent program to transition the root tip from one that is actively growing to one that terminally differentiates following Al treatment. Some genes that were inferred to be SOG1 independent were also evaluated by qPCR to test if Al response involves the DDR and additional pathways that mediate distinct processes such as Al-dependent organic acid release. In total the gene targets chosen for qPCR were selected based on being differentially expressed under the tested conditions or of interest for putative roles in the Al response. In addition genes known to be involved in DNA repair as part the DNA damage response or expressed from biochemical pathways related to Al toxicity were also included in the qPCR target set.

Table of select differentially expressed genes identified from RNAseq (condensed)							
Gene ID	Primary Gene Symbol	Col-0 FC	Col-0 FDR	*sog1-7* FC	sog-1 *FDR*	*als3-1* FC	*als3-1* FDR
AT5G07620	Unknown	2.28	0.002	0.20	1.000	4.81	0.000
AT1G04450	ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN 3 (RIC3)	3.10	0.454	1.15	0.746	4.71	0.001
AT1G13980	GNOM (GN)	2.33	0.109	0.11	1.000	2.77	0.001
AT1G13330	ARABIDOPSIS HOP2 HOMOLOG (AHP2)	1.77	0.109	0.70	1.000	1.54	0.001
AT1G60500	DYNAMIN RELATED PROTEIN 4C (DRP4C)	2.42	0.321	0.66	0.852	3.36	0.000
AT3G09670	Unknown (Tudor Protein Super Family)	3.08	0.000	2.90	0.000	3.02	0.000
AT5G22890	SENSITIVE TO PROTON RHIZOTOXICITY 2 (STOP2)	2.64	0.000	1.79	0.002	1.91	0.000
AT3G05820	INVERTASE H (INVH)	-0.39	0.738	-0.53	0.435	-1.12	0.003
AT5G59440	ZEUS1	-2.19	0.410	-0.25	1.000	-4.62	0.000
AT5G53200	TRIPTYCHON (TRY)	-1.05	0.738	1.41	0.744	4.49	0.001

Table 1: Condensed List of Gene targets
List of differentially expressed genes further explored for analysis in the potential role of the response to aluminum toxicity.

Multiple gene targets were explored based on a combination of the results differential expression, and literature, if possible, to support a role the response to Al toxicity, such as *STOP2* or relation to the DNA damage response regulated by SOG1 such as *HOP2*. These gene targets include *At1g04450* (*RIC3*), *At1g13980* (*GNOM*), *At1g13330* (*HOP2)*, *At1g60500* (*DRP4C*), *At5g22890* (*STOP2*), and *At3g05820* (*INVH*) and unknown proteins, see Table 1 for the condensed list. The unknown proteins were included based on protein domains such as kinase domains in the case of *AT5G07620*, or experimental data such response to abiotic stress from electronic fluorescent pictograph (eFP) data shown in Figure 11 [50], thus providing outside evidence in support of the gene playing a role in the response to Al exposure. While some of these genes did not fit the pattern explained previously to be considered SOG1 regulated, these genes were seen as not only controls to validate the RNAseq results, but also exploring the findings that could help research expand to other areas of the response to Al toxicity.SOG1 Dependent Genes

RIC3 previously was characterized as primarily being expressed in the pollen tubes of Arabidopsis [51] although expression in other tissues such as the root was not investigated. *RIC3* plays a role in the polar cell expansion via actin recycling in conjunction with available calcium through pathways with Rho-GTPase 1 (ROP1) to either disassemble F-actin and lead to cell growth or lead to its inhibition [52] [53]. While the previous research has demonstrated this process it's very probable that the same process for cell expansion could be occurring in other tissues of the plant in similar pathways. The support for this hypothesis comes from the RNAseq and qPCR data show that in response to Al expression of *RIC3* increases, in both wild type and *als3-1*.

However 3.5 day old *Arabidopsis* seedlings have not developed this type of tissue. This indicates that *RIC3* may be found in more tissues than previously reported. Additionally there is evidence to suggest that Rho-guanine nucleotide exchange factors (RhoGEFs) a family that includes factors, such as *RIC3*, interact physically and with receptor like kinase proteins [54] with one possibility that could include *At5g07620* which was identified in the RNAseq analysis and will be discussed later on. This type of interaction with RLKs has been speculated to have coevolved for form a plant specific signalling path way that responds to extracellular signals, which could include things such as ROS, or Al^{3+} [55]. There were was another GEF superfamily protein that was identified in the analysis, *GNOM.*

In previous studies of *GNOM* the phenotype is identified as also resembling those that occur when interfering with auxin transport [56]. Using interfering RNA to knockdown the expression of *GNOM*, which is expressed in the root and results in the short root phenotype as part of the disruption of the polar transport of auxin [57]. *GNOM* plays a role in auxin polarization and transport as an ARF-GEF which is a class of guanine nucleotide exchange factors, is responsible for the regulation of the vesicle transport of *PINFORMED 1* (*PIN1*), an auxin efflux carrier, whos localization leads to auxin polarization [58] [56]. In both wild type and *als3-1*, Al treatment gives a 2 fold induction compared to control conditions.

In contrast, for *sog1-7* and the suppressor mutants, Al treatment results in no induction. This suggests the possibility that if the *als3-1* phenotype which resembles *gnom* could mean that the Al response is in some way may also interfering with the auxin transport or polarization.

AtHOP2 encodes a protein that functions to aid in DSB repair via inter homolog strand bias, helping to create and resolve the Holliday junction used as part of homologous recombination [59]. The majority of research done on *AtHOP2* involves its role in meiosis in conjunction with MEIOTIC NUCLEAR DIVISION PROTEIN 1 (MND1) to form the HOP2/MND1 complex. AtHOP2 and MND1 have been determined from previous research to be required for homologous recombination, with notes that failure of this complex can lead to incorrect homology or fragmentation of the chromosomes [59,60]. At*HOP2* was also identified to inducible with ionizing radiation, but more interestingly At*HOP2* was found to be induced in an *ATR* dependent manner and not by *ATM* [61].

This lends even more support as it fits the previously theorized model that damage from Al toxicity generates a response that is *ATR* and *SOG1* dependent. At*HOP2* is found in several different eukaryote species and plays a role in both meiosis and DSB repair [59]. In support of the RNAseq results in which expression of At*HOP2* is Al-inducible in a *SOG1* dependent, *HOP2* expression increases in wild type by 1.7 fold and by almost 8 fold in *als3-1*, whereas expression is effectively unchanged in *sog1-7*. A role for *AtHOP2* in response to Al would be consistent with the argument that Al acts as a DNA damage agent, with it being predicted that a loss-of-function *athop2* mutant would be likely hypersensitive to Al due to failure to repair Al-dependent DNA damage. Through the use of an online database for gene expression which can be found at

University of Toronto site (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi) [50], further support of AtHOP2 playing a role in a DNA damage response comes from the eFP browser for the gene which shows a massive increase in expression with genotoxic agents but no change in expression with oxidative stress.

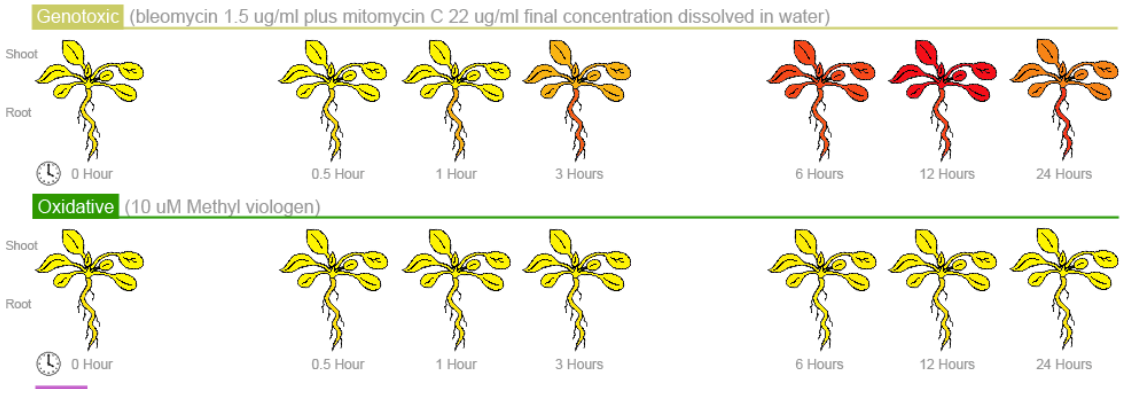


Figure 11: eFP results for At1g13330
Results for AtHOP2 indicate that protein is active past just meiosis and responds to genotoxic stress, as shown in the first row. However it does not change its expression from a purely oxidative stress shown in the second row [50]. This provides additional support to the theory that DNA damage is occurring as a result of Al^{3+} exposure when correlating the expression of AtHOP2 with the different stress conditions.

DRP4C (*DYNAMIN RELATED PROTEIN 4C*) encodes a GTPase, which is an enzyme responsible for interacting with G-Proteins to hydrolyze GTP to GDP in order to control the activity of G-Proteins. The DRP4 class of proteins is related to mammalian antiviral response in animals, however DRP4C function has yet to be characterized [62]. Dynamins are found in all eukaryotes and function to facilitate vesicle budding from the plasma membrane [62] as well many different members of the Dynamin family are responsible for different fusions or divisions in various membrane systems [63]. From the RNAseq results DRP4C has a 2 fold increase in wild type and 3 fold increase in *als3-1*, and follows the pattern of being *SOG1* regulated with a fold change of almost zero.

However further study would be required to determine if *DRP4C* is directly regulated via sog1. From other results with genes like *GNOM* or *RIC3* that have to do with vesicle transport, its possible this gene could also play a similar role, though not well characterized in the literature as to what that might be.

ZEUS1 (*AT5G59440)* encodes a THYMIDYLATE KINASE with multiple isoforms that depending on the isoform goes to different parts of the cell. *ZEUS1* synthesizes dTDP and is involved in the regulation of DNA replication, specifically during the G1/S phase *ZEUS1* is at its peak expression [64]. Interestingly, in contrast to other identified SOG1-regulated genes, *ZEUS1* is SOG1 regulated in a negative manner. As seen in the RNAseq results, when exposed to Al, transcription of ZEUS is repressed 2-fold in wild type and 4-fold in *asl3-1*, but is essentially unchanged in *sog1-7*.

At5g07620 is a gene that encodes a protein that is predicted to be part of the receptor-like kinase (RLK) family, currently the protein is listed as having an unknown function and there are no published reports demonstrating its role in plants. Through further testing and validation via qPCR this gene was confirmed to have the expression pattern shown in the RNAseq. The biological process is described as being involved in protein phosphorylation. The cellular component is that of the component of the plasma membrane. While the molecular function provides ATP binding and with (protein) kinase activity.

Since the phenotype of Al-treated *als3-1* is largely related to SOG1-dependent endoreduplication of the root, it could be argued that expression of factors that are key to this terminal differentiation should be found in this tissue. Using the eFP browser [50] it was found that along with other tissues including the shoot and leaves, *At5g07620* is in

fact expressed in the root. Use of the bar eFP browser provided by University of Toronto, which shows expression of genes of interest under different conditions, it was found that following exposure to various conditions that could be considered analogous to Al treatment including oxidative and genotoxic stresses, *At5g07620* shows increased expression. Combined these results suggest that *At5g07620* is expressed in the root in a manner dependent on genotoxic stress, which would be expected for a factor that is induced in response to Al exposure.

qPCR analysis indicates that Al responsive *At5g07620* expression is both ATR- and SOG1-dependent, indicating that this may be a key downstream component of the ATR-mediated stoppage of root growth following Al treatment. As shown in Figure 13, qPCR analysis with *At5g07620* supports previous findings of the model in which loss-of-function mutations in either *ATR* or *SOG1* result in full suppression of the als3-1 Al hypersensitivity. As with this full suppression and contrary to Al treated *als3-1*, *At5g07620* is not upregulated following Al treatment in either *atr-4;als3-1* or *sog1-7;als3-1*, thus suggesting a key role for this factor in terminal differentiation following Al treatment.

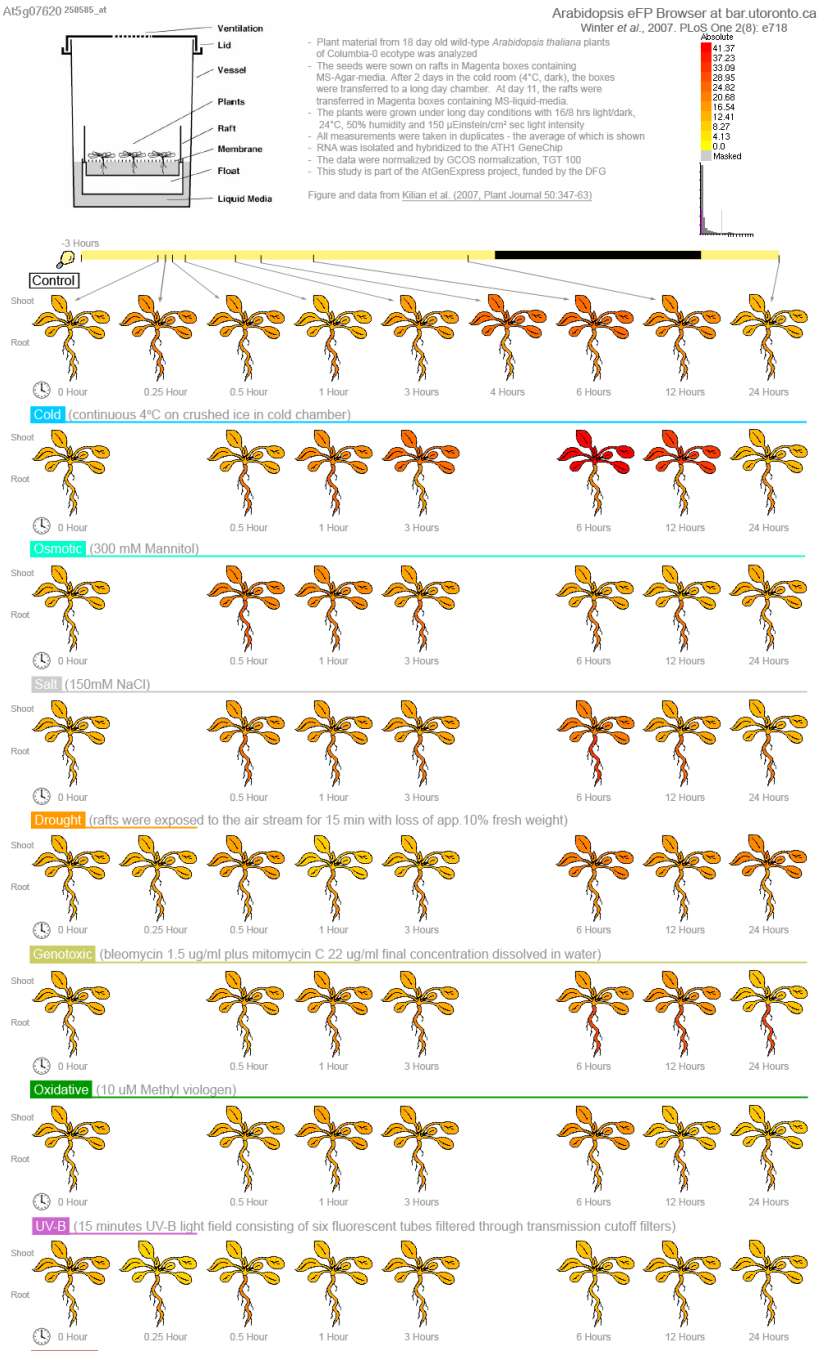


Figure 12: eFP results for At5G07620
Figure from University of Toronto eFP website for arabidopsis. Using the gene At5G07620, this browser allows us to speculate based on the results of the eFP as to the changes of this gene to multiple stresses including but not limited to genotoxic and oxidative stress. (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi) [50]

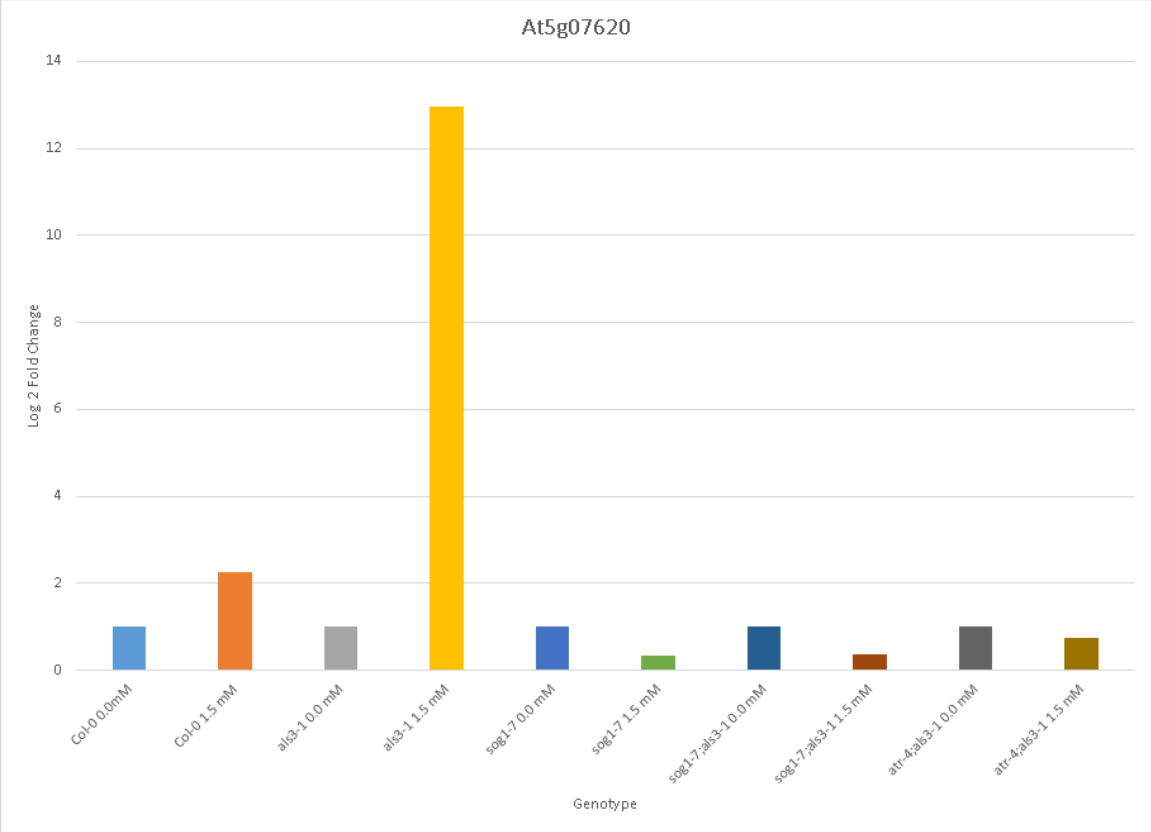


Figure 13: qPCR Analysis of At5g07620
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

SOG1 Independent Genes

The Tudor protein super family, of which *At3g09670* is a member, is generally described as having roles in both cell growth and differentiation. Contrary to other genes of interest that were identified from the RNAseq analysis, this particular factor was found to be upregulated by Al similarly for Col-0 wt, *als3-1* and *sog1-7*, thus suggesting that expression of this factor is SOG1-independent and likely is not under control of the DDR. This is consistent with SOG1-controlled gene expression being only one aspect of Al response.

STOP2 (*AT5G22890)*, which is a zinc finger like protein that regulates genes relating to pH and other stresses, with several of these also regulated by its homolog, STOP1 [23]. Where these two proteins differ is that STOP2 is more responsible for those genes that are largely specific to Al toxicity such as *AtMATE* and *ALS3* [23]. RNAseq analysis shows uniform increase in the fold change across all genotypes, with the highest being wild type of 2.6 fold change, and 1.9, and 1.7 fold change in *sog1-7* and *als3-1*. As with *At3g09670*, *STOP2* also appears to be Al-inducible in a SOG1-independent manner thus suggesting it is part of a pathway that is not related to the DDR.

INVH, or ALKALINE/NEURTRAL INVERTASE H (*AT3G05820)* also appears to be SOG1 independent in a manner different than STOP2, and is seemingly interesting since it plays a general role in overall plant growth and development [65]. RNAseq analysis shows the expression being down-regulated, less than one fold change

between the treated and untreated samples for wild type and *sog1-7* but seems to be more down regulated in *als3-1*. It is also interesting due to the role is plays specifically in the plant response to stress, including having a role in ROS formation. A loss-of-function mutant for this gene results in roots that do not generate ROS [65,66]. Since the RNAseq data shows a greater reduction in expression of INVH in *sog1-7* compared to the other tested lines, it is possible that dampened levels of INVH might reduce the levels of Al-dependent ROS in *sog1-7* and result in the observed Al tolerance.

TRY (*AT5G53200*), encodes TRIPTYCHON, which functions as a transcription factor involved in the inhibition of lateral root hair growth in Arabidopsis [67]. *TRY* expression is novel with regard to a relationship to Al and SOG1 in that it shows only limited changes for both Col-0 wt and *als3-1* following Al exposure yet is highly upregulated for Al-treated *sog1-7*. This gene is interesting, due to the very different expression in each genotype, its being repressed in Col-0 down 1 fold, up 1 fold in *sog1-7*, and very highly expressed in *als3-1* with 4 fold induction. Which shows that it is independent of SOG1 but also gene that could be playing a role in the very different phenotypes we see between Col-0 and *als3-1* when exposed to Al^{3+}.

Network Analysis

Since these DEGs were identified and curated, there was a desire to see if any of them interacted with one another in a yet undetermined pathway. To explore this a network analysis of the protein interactome was performed to determine which factors that interact with our identified Al-inducible gene products might also be involved down stream as part of the response to Al toxicity in the plant. The ENSEMBLE database was

used for this analysis (IntAct) [68] which is a database of protein interactions that come from many different sources. It allows a user to submit genes of interest and it returns a spreadsheet of interacting proteins. From there programs like cytoscape can be used for visualization [69]. Many of these genes, especially unknown and characterized ones, seem to only have one or two nodes of connection while others such as GNOM, TRY or STOP2 show many nodes that could be potential targets for further investigation.

Genes that encode transcription factors such as TRY, GATA Zinc Fingers and STOP2 are expected to have many edges or connections in the interactome. Multiple edges (denoting interactions) leading to the same node in the interactome mean that multiples sources have confirmed that protein-protein interaction, thus providing more certainty regarding the relationships. It is also relevant to point out that it is also still limited by the what studies have been done by the scientific community, for example those proteins whose function has not yet been characterized may not show up in the results unless it was in an assay where it was pulled out with some other known protein. Of note, when SOG1 was included in this analysis from the IntAct database there was no interaction data available.

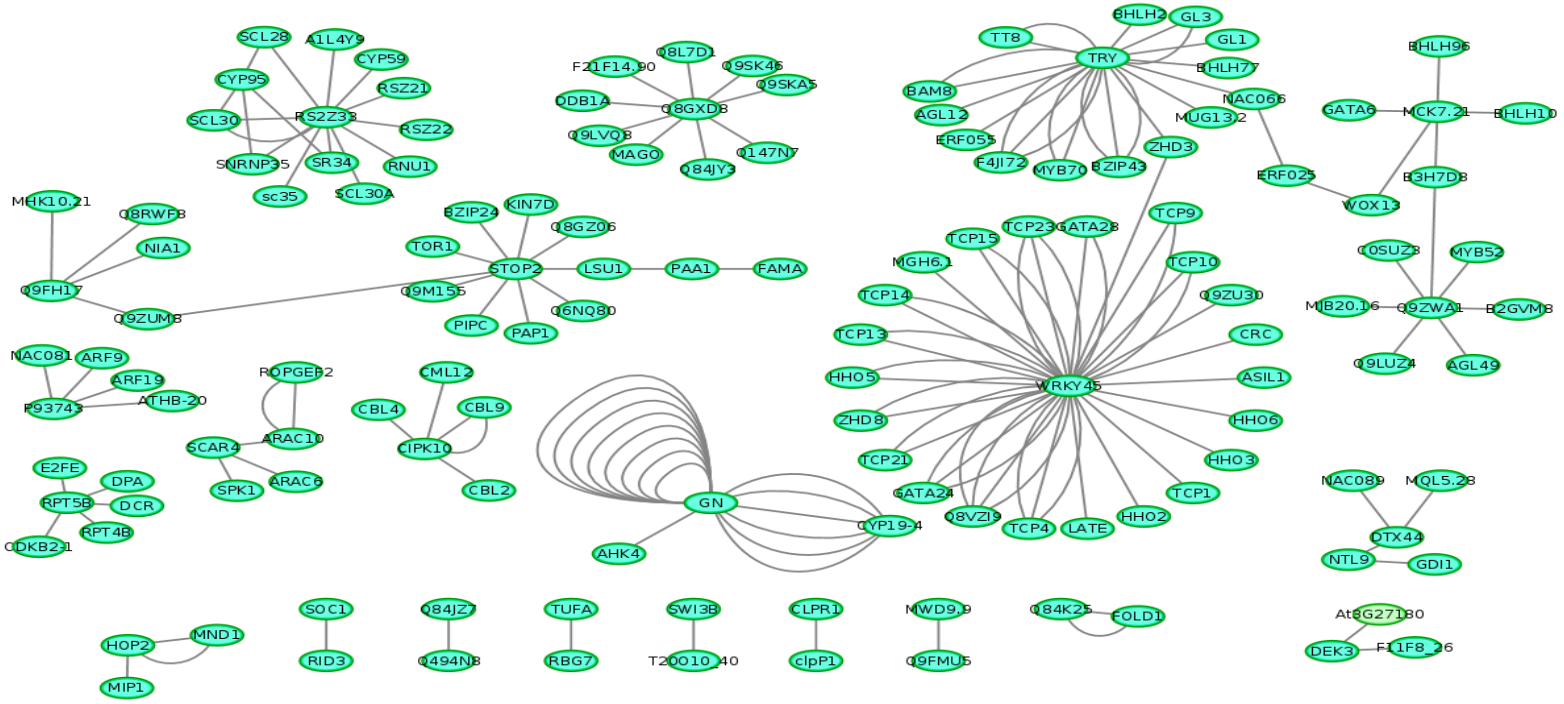


Figure 14: Cytoscape Interactome Plot
Using the DEGs that were identified using the criteria of ≥2 fold change and ≤1% FDR were then queried through ENSEMBLEs IntAct database to determine gene production interactions. Those interactions are what are displayed here using Cytoscape.

qPCR Validation of Selected Gene Targets

For those qPCR targets the following expression profiles followed:

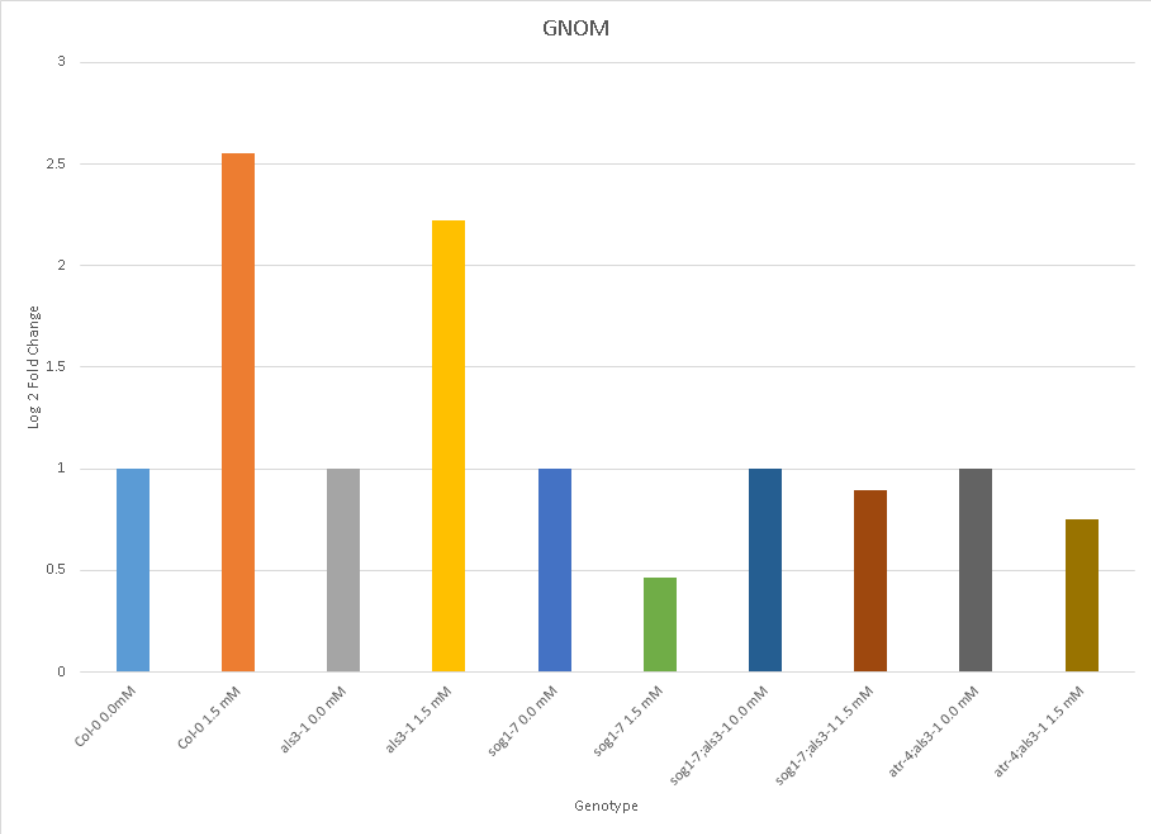


Figure 15: qPCR Analysis of GNOM
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

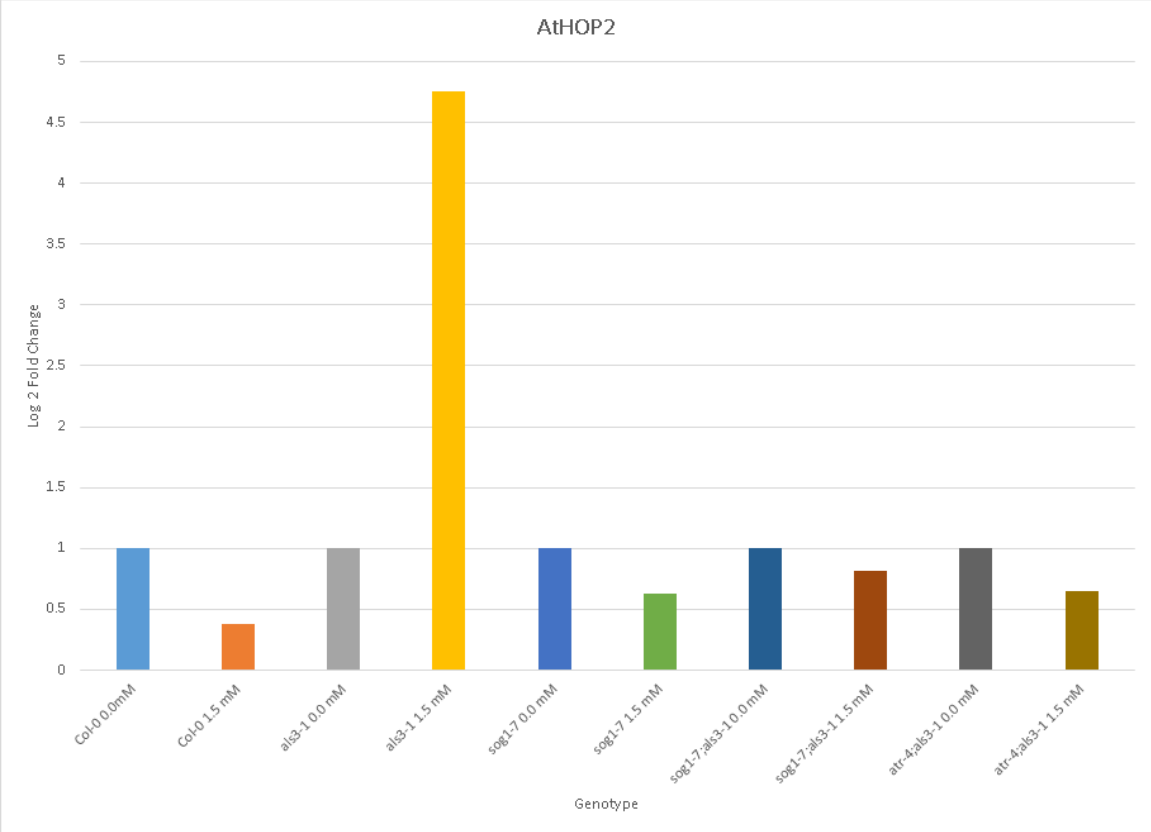


Figure 16: qPCR Analysis of AtHOP2
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

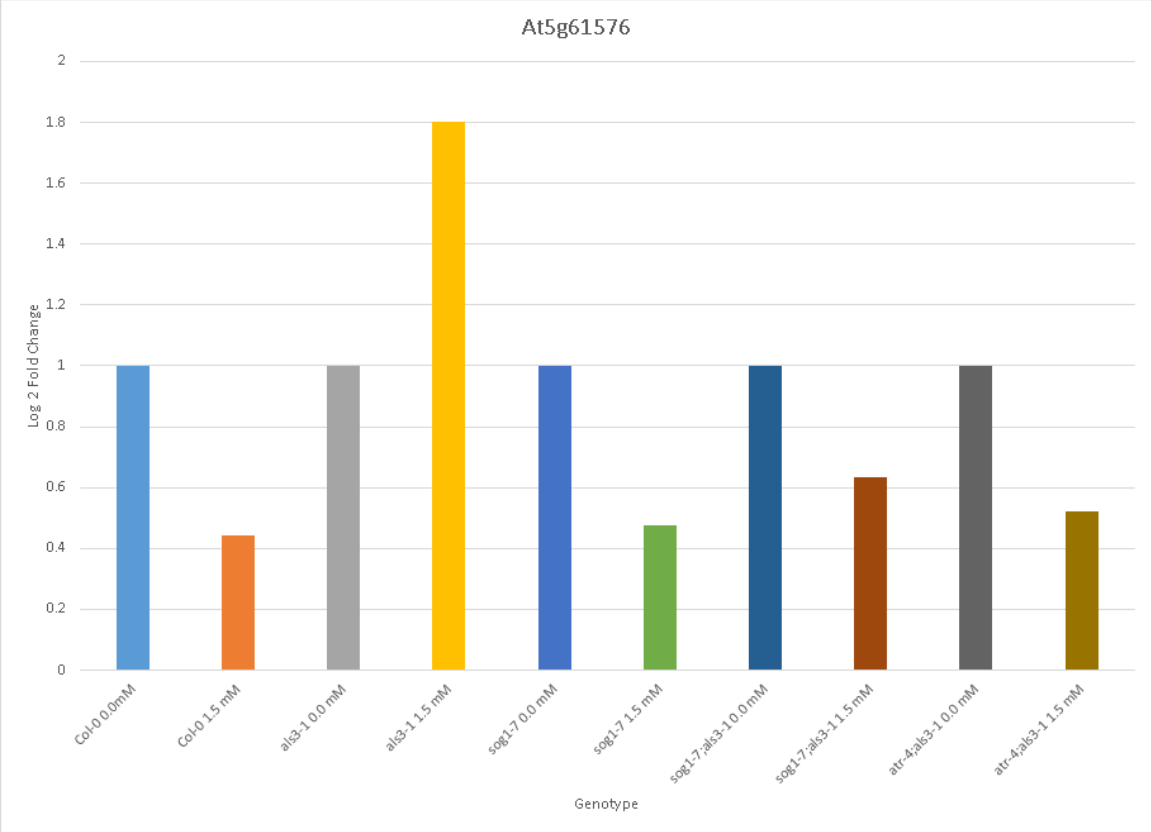


Figure 17: qPCR Analysis of At5g61576
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

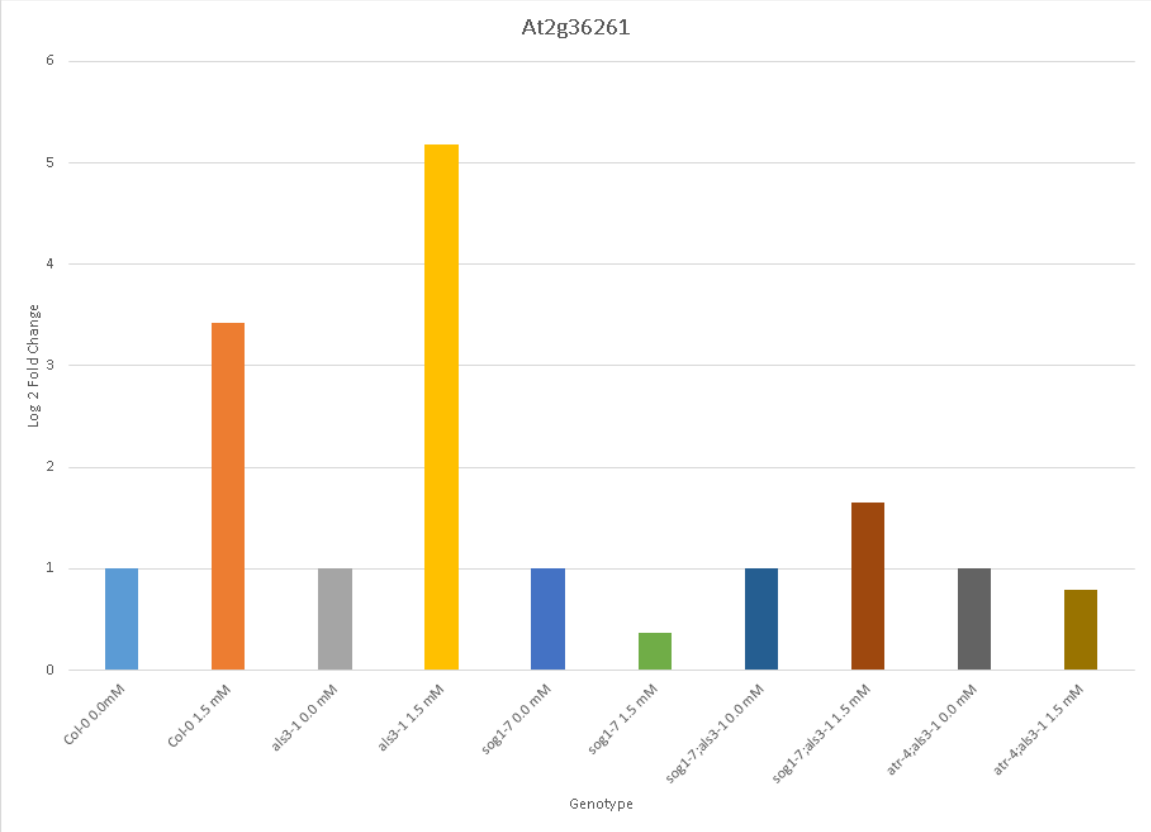


Figure 18: qPCR Analysis of At2g36261
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

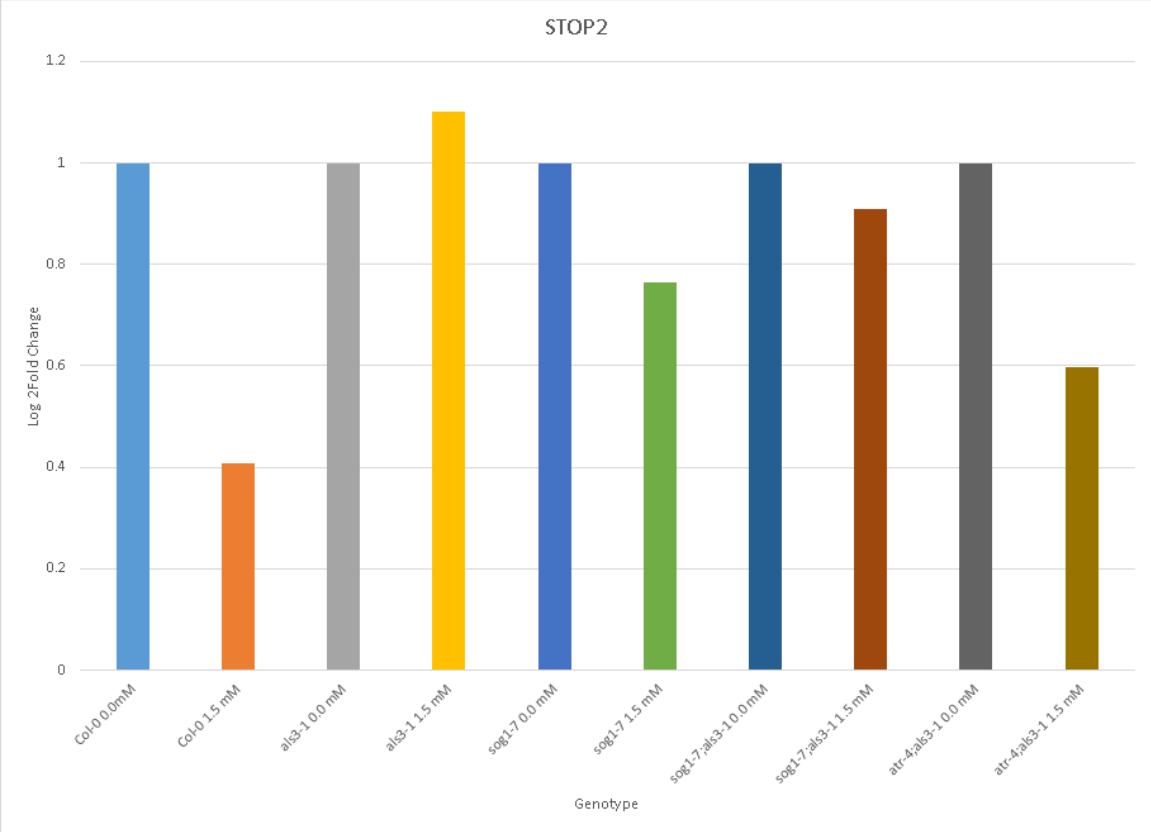


Figure 19: qPCR Analysis of STOP2
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

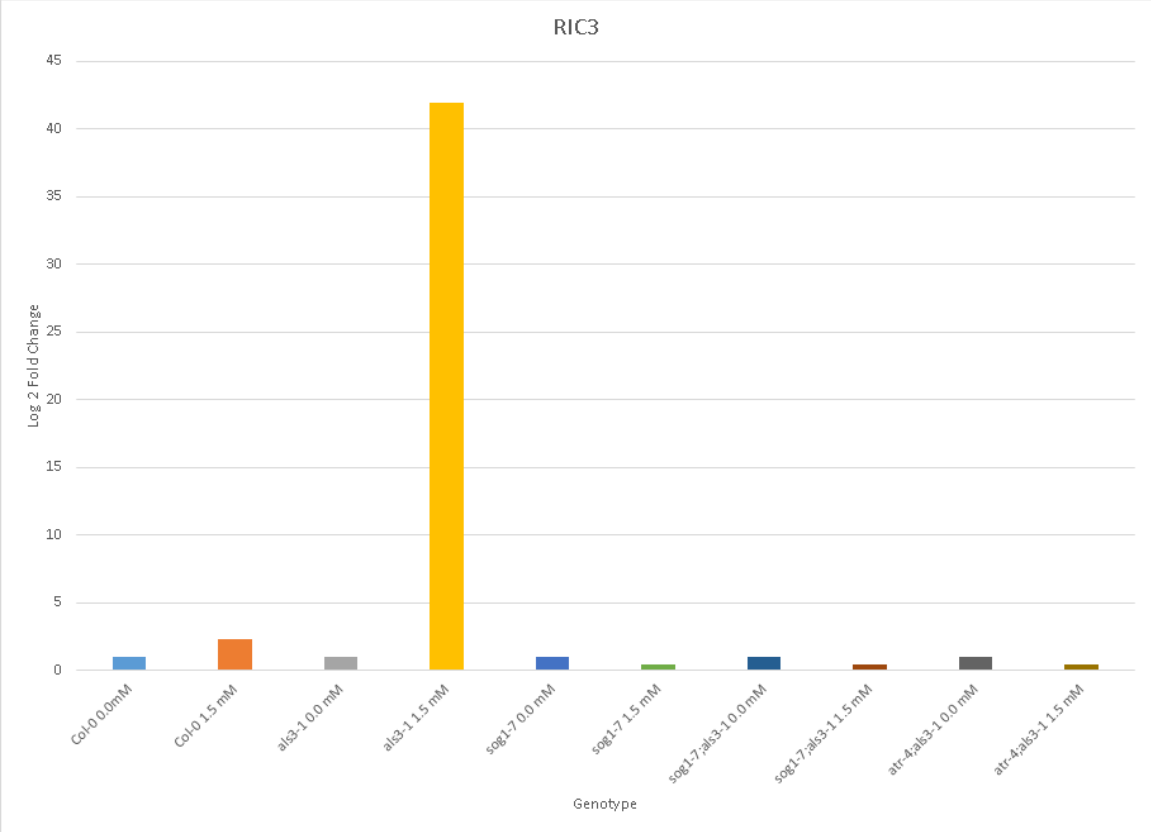


Figure 20: qPCR Analysis of RIC3
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

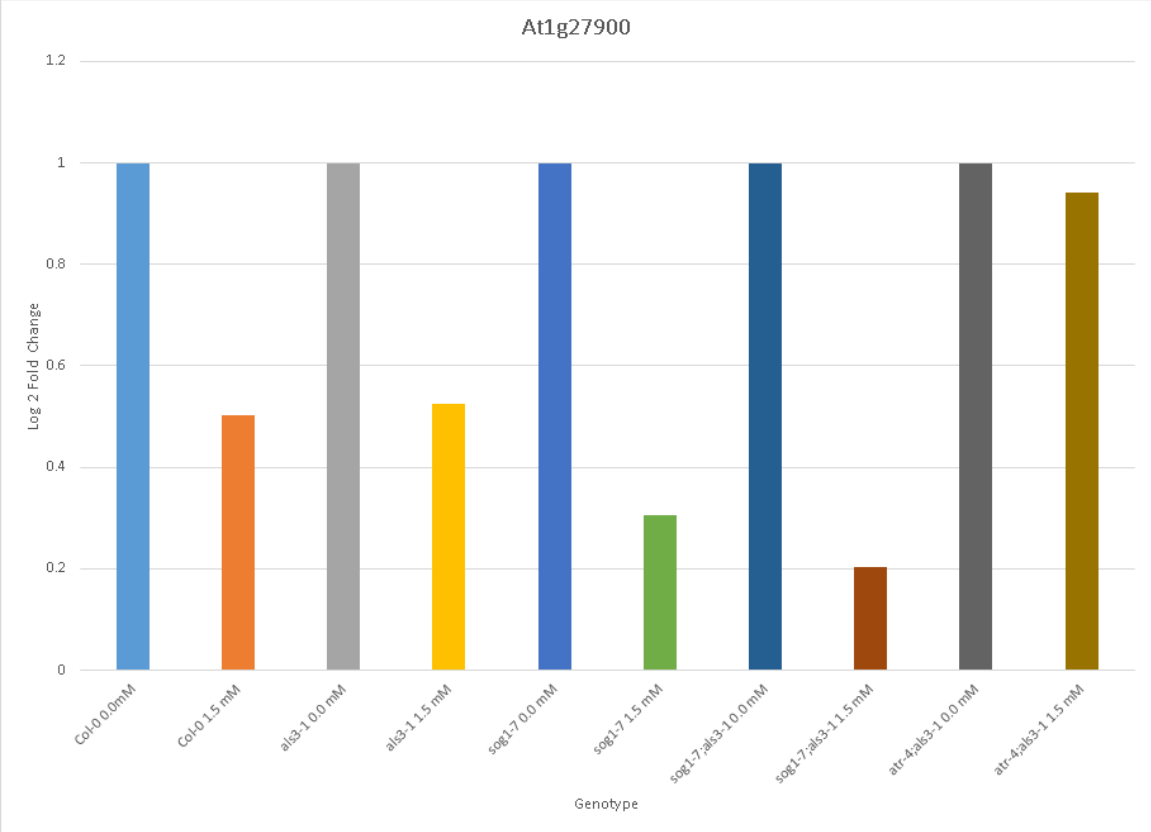


Figure 21: qPCR Analysis of At1g27900
The gene expression is normalized to a control gene *Ubiquitin Carrier Protein 1* (*UBC1*) after which the relative expression is compared to that of the same genotype on the control media (0.0 mM) which is set to 1 and then the changes of the treated genotype are determined from the difference. The qPCR in this manner is then comparable to the results of the RNAseq. The genotypes used are Col-0, *als3-1*, *sog1-7*, which represent the validation of the RNAseq and *sog1-7;als3-1*, *atr-4;als3-1* to expand the results and confirm if the results fit the current working model of being ATR activated and SOG1 dependent.

Discussion

RNAseq is an experimental approach that allows the whole transcriptome to be surveyed and is therefore a very powerful tool that provides a large amount of data to identify genes that are transcriptionally regulated in relation to a particular biological

process. For the experimental setup for RNAseq analysis to identify SOG1-dependent Al inducible genes, gel media that was soaked with either 0.0 mM or 1.5 mM AlCl3 (pH 4.2) was used to provide an Al toxic environment equivalent to the relatively high levels of Al found in soils, which was approximately equal to 100 µM. After growth for approximately 3.5 days, the seedlings were harvested. This time point was chosen since previous work had shown that this represented the approximate time following Al exposure at which growing root tips began the developmental process of terminal differentiation [26].

For this analysis, Col-0 wt was compared to the mutants *sog1-7*, which is considered to be Al tolerant, and *als3-1*, which is Al hypersensitive. Inclusion of these mutant genotypes was done in order to define a sub-group of Al responsive genes that are SOG1 regulated to potentially identify key factors responsible for Al-dependent terminal differentiation. *als3-1* was utilized to provide confirmation that those genes were clearly induced, since Al-regulated genes would be expected to be modestly expressed in Al-treated wt but highly expressed in Al-treated *als3-1*. *sog1-7* was included to identify genes that are SOG1 regulated, since such genes should be Al-inducible in Col-0 wt, hyper induced in *als3-1* but show how increase in expression in Al-treated *sog1-7*. This is in contrast to other Al-inducible genes that likely have increases in expression independent of SOG1. Such an approach should allow for the identification of genes that are key to Al-dependent terminal differentiation since this has been shown to require functional SOG1 and its key regulator, the master cell cycle checkpoint ATR

SOG1 Regulated Genes

If this model is correct, then mutational loss of these genes should lead to Al

tolerance and failure to initiate terminal differentiation and endoreduplication. Initial analysis of a loss-of-function mutant for *At5g07620* has so far not revealed the role of this factor in Al response, although the substantial increase in expression of this gene following Al exposure suggests it may be important. Using the gene model of the transcript *At5g07620* is predicted to encode a receptor like kinase protein. Predictive analysis of the protein domains encoded by *At5g07620* indicates that it lacks several hallmarks that would be required for kinase activity, suggesting that while it might be classified as an RLK, it likely is not a functional kinase. Additionally, the predicted protein encoded by *At5g07620* does not appear have an extracellular domain, suggesting that it does not have the capability to receive signals from outside of the cell. Based on these analyses, it is possible that the role of this protein may be related to inhibition of other RLKs by preventing them from binding to their targets. One example of this could be in the role of hormone signalling with proteins such as that encoded by *At5g07620* possibly short circuiting the signaling pathway regulate the hormone response.

Plant hormones are responsible for various plant functions including but limited to: reproduction, growth, and development. Since the expression of *At5g07620* is increased in relation to Al treatment and is greatest in the Al hypersensitive *als3-1* background, it is arguable that in some way this factor are key to Al dependent remodeling of the root tip. Further analysis will be required to test this hypothesis, which is beyond the scope of the presented work.

Other genes that were identified from this RNAseq include promising candidates such as *GNOM*, *HOP2*, and *RIC3*. These proteins have been characterized as being

involved in processes such as hormone regulated morphological changes [57], DNA repair pathways [59] and plant response to disease [51]. It is also not surprising that in the results there are proteins related to Al resistance such as MATE. This is expected that these would increase in expression due to the toxicity. However they do provide a good positive control for genes that are expected in the analysis. It is of note that using 2 fold change, with a false discovery rate of one percent, is some strict in order to be more confident in the results of the analysis. It is possible that this analysis could be rerun with a FDR of five percent and have an even larger pool or potential targets to choose from.

GNOM in particular seems promising in terms of the morphological changes that occur as a result of loss of *GNOM* function, with mutants having a short root with a swollen tip [57]. This resembles the phenotype seen in *als3-1* when exposed to Al although it is not clear why upregulation of *GNOM* following Al exposure would give the same type of phenotype as a loss-of-function *gnom* mutant. Possibly Al toxicity not only causes hyperexpression of GNOM but also alters its function. Testing this hypothesis will be challenging since *gnom* mutants are adult sterile and as mentioned, loss of *gnom* likely will not suppress the *als3-1* Al dependent endoreduplication. Certainly though, further investigation of the role of GNOM in Al response is warranted based on the RNAseq data. However as this gene has been characterized to have a similar short swollen root phenotype and seeing it in the RNAseq results especially induced for *als3-1* lends to the hypothesis that GNOM is involved when these plants reach the point of exposure where the roots undergo terminal differentiation.

Interestingly, GNOM physically interacts with with ARABIDOPSIS HISTIDINE KINASE 4 (AHK4), which is responsible for binding to cytokinins and transporting them

across the plasma membrane. Cytokinins are a plant specific hormone that play a prominent role in the cell cycle, specifically cell growth, growing larger and faster[68]. Roles of cytokinin are tissue specific, as it has opposing effects in the root compared to the shoot [70]. Additionally, cytokinins work synergistically with other hormones such as auxin to regulate the root meristem. For example, in response to Al exposure cytokinins mediate the auxin signaling [71]. This provides further suggestive evidence for a role for GNOM in how plants respond to Al.

RIC3 is part of the same superfamily as GNOM, with each of them considered to be ROP (Rho GTPase) interacting proteins. While GNOM has been characterized by its short root phenotype, RIC3 has mainly been characterized in terms of its role in pollen tube growth [51]. There currently is no data regarding where else in the organism it might be expressed or its molecular function which in terms of the GO information is simply listed as protein binding due to the targeted nature of the study. Since it controls cell expansion in pollen tubes, it arguably could perform the same role in the root, which would be consistent with Al altering cell expansion in conjunction with endoreduplication. Brefeldin A disrupts transport and recycling of GTP, which consequently disrupts G-Protein function.

It is also interesting to note that DRP4C also has a role in GDP-GTP exchange, suggesting that it too might also be part of a mechanism comprising RIC3 and GNOM. qPCR results show that expression of *RIC3* is upregulated substantially in response to aluminum specifically in *als3-1*. This could be related to an important role for RIC3 in promoting some of the *als3-1* specific phenotypic responses to Al that have been reported. Based on previous studies done with the pollen tubes it could be a potential

factor that plays a major role in the terminal differentiation of the root or enlarged cell size characterized seen for Al treated als3 loss-of-function mutants [72].

Additionally ROP-GEF (ROP-Guanine Nucleotide Exchange Factor) according to the network analysis shows 2 ROPs that interact with ROP-GEF2. These ROPs are important as they serve as developmental signals that bind to GTP, two that were mentioned previously were included in RIC3 pathways with the cytokinins and GNOM with the auxin polarization. In order to activate these developmental signals GEFs are required [73], which could provide additional evidence as to what factors could be involved in the stoppage of root growth if expression of these factors is being reduced in *als3-1* following Al treatment. It could be speculated that modifying a plant to maintain expression of these genes in the presence of Al could block manifestation of the *als3-1* Al hypersensitivity phenotype.

HOP2 has a demonstrated role in homologous recombination in various eukaryotes due to its role in DNA pairing [74]. The observation that At*HOP2* expression is SOG1-regulated in an Al-dependent manner provides suggestive evidence of what type of DNA damage might occur as a result of Al interfering with DNA. HOP2 specifically is responsible for interstrand homologous binding used during repair as part of meiosis as part of a complex with DMC-1, This is complex is responsible for the proper binding, recombination and segregation of sister chromosomes during meiosis [59,75]. This is important because the genomic consequences of Al exposure are not well understood and attribution of a particular repair pathway to Al toxicity will give insight into the consequences of Al on DNA. Previous research show factor like ATR and SOG1 are involved in the response, and there are phosphorylated ɣ-H2AX which signifies the

presence of double strand breaks. As part of the DSMC-1 complex, HOP2 works in conjunction with MND1, DMC-1 and RAD51 at least in meiosis to repair these breaks, though the possibility exists that they could be active during other stages of life and growth in Arabidopsis [59].

MND1 has been characterized as being part of DSB repair during meiosis [59]. Since At*HOP2* is induced in an Al dependent manner in young seedlings that are not engaged in meiosis, it is possible that these proteins play a role in DSB repair following Al exposure. This suggests that the DSMC-1 complex may have a role outside of meiosis or that HOP2 could be involved in another pathway independent of the DSMC1 complex that repairs DNA damage resulting from Al exposure.

There is a reasonable likelihood that homologous recombination is involved at least to some extent following Al exposure. HR as a repair mechanism is for the most part error free, due to using the sister chromosome to repair any damage with an undamaged template [30]. HR is relatively limited temporally as it can only occur at distinct points during mitosis following DNA replication and generation of sister chromatids. Consequently, mechanisms other than HR may come into play following Al-dependent DNA damage and those will be discussed in the following chapter.

SOG1 Independent Genes

Arabidopsis STOP2 is a zinc finger protein that is a homolog of STOP1, both of which have been found to lead to transcription of genes that are expressed under low pH conditions and following exposure to Al [23]. This provides additional evidence in support of our RNAseq results as a gene that is responsive to Al^{3+} it is a gene that we would be

expecting to see increased expression especially in the sensitive mutant *als3-1*. Based on the finding of [23], loss-of-function *stop2* plants grown on low pH show a root length comparable to WT, but are substantially inhibited compared to wt when exposed 2 µM Al. This is suggestive that STOP2 function is specifically related to Al response rather than the low pH. In loss-of-function *stop1* mutants with functional STOP2 to determine the role of STOP2 it turns out that with low pH a total of 11 out of the 81 STOP1 genes are are still induced, 3 of which are specific to only low pH, while 45 genes are differentially expressed post exposure to AL, 8 of which are in common with the low pH leaving 37 Al specific genes that appear to be regulated by STOP2 [23] these genes are however not further discussed in the paper and cannot be compared to our analysis. This would seem consistent with the previous research showing that STOP2 is not involved with the DDR, but instead is part of another path way that is part of disease and stress response in the plant. Based on the network analysis of the transcriptome, there are many other proteins that STOP2 interacts with, most directly and some as secondary targets through AT2G19650 (Q9ZUM8) and RESPONSE TO LOW SULFUR1 (LSU1) (Figure 14).

Invertase H (INVH) is differentially expressed among all three genotypes in relation to Al with a fold change of -0.39 in wild type, -0.56 in *sog1-7* and -1.11 in *als3-1*, thus demonstrating its Al dependent expression is independent of SOG1. INVH's role in the plant is to generate reactive oxygen species in response to DNA alkylation [65]. A possible role for INVH could be to create ROS that can bind to cations [65], in this case Al^{3+}, to prevent additional Al dependent DNA damage such as in the form of DNA alkylation [41]. INVH when knocked out leads to a phenotype of a severely shortened

shoot although it is also expressed in the root, which is logical based on its function in ROS response and the root of the plant being the primary source of exposure to the cationic stress. This protein when knocked out leads to no ROS being produced in the root. If ROS are important for binding to Al to oppose toxicity, the role of INVH could be very crucial to Al detoxification in the plant. However, ROS can lead to DNA alkylation that in turn can lead to replication fork stalls as well as replication fork collapses [41]. While this type of damage can be repaired, often this can occur through error prone repair pathways that could lead to unintended DNA damage.

ZEUS1 regulates the replication of DNA through its role as thymidylate kinase, which are responsible for acting as the catalyzing enzyme for dTDP production, which is the backbone of one of four nucleotides (dTTP), with necessary implications for synthesis of new DNA [64]. It has also been characterized as being regulated by the G1/S checkpoint of the cell cycle [64]. ZEUS1 expression in a loss-of-function *sog1* is unchanged in relation to Al (-0.25 fold change), as opposed to being significantly downregulated in wild type (-2.19 fold change) and especially in *als3-1* (-4.62 fold change).

Its down regulation provides insight to a phenomenon first noticed as part of the *asl3-1* phenotype in which the plant undergoes endoreduplication, a process under which the organism begins to create many copies of its DNA without cytokinesis. ZEUS1 was hypothesized to play a role in cell division [64] the down regulation of such a factor could be playing a large part in the process that leads to the hyper enlarged cells with many copies of the DNA. Since ZEUS1 is regulated by SOG1 this could signify that as part of the DDR SOG1 activation leads to inhibition of ZEUS1 following exposure to Al to

prevent cell division in order to prevent these damage cells from growing out of control and possibly compromising the plant as a whole.

While qPCR is useful as a validation of the RNAseq data, it is not necessarily comparable to RNAseq since each method is analyzed in a different manner. In the case of RNAseq, the method deals with absolute counts and then uses statistics to determine if that gene is differentially expressed. qPCR is instead based on primers designed to amplify specifically a gene target, with the fold change being determined based on relative expression compared to a selected reference gene. Previous qPCR experiments relating to SOG1-dependent changes in gene expression following Al exposure, *ELONGATION FACTOR 1* (*EF1a*) *At5g10630* was used as a control for determination of fold changes for Al responsive genes [26]. RNAseq analysis using the same growth conditions, found that *EF1a* is actually differential expressed after Al exposure with values of approximately -0.9 fold change in Col-0, -2.1 in *sog1-7* and -3.6 in *als3-1*, which could explain why the RNAseq and previous qPCR results do not necessarily correlate. Where as in this analysis UBIQUITIN CARRIER PROTEIN 1 (UBC1) was used which was also checked against the RNAseq data and did not show any differential gene expression.

The RNAseq analysis combined with the network analysis of the interactome provides a broader picture of what other genes and pathways might be related to the Al toxicity response thus providing directions for further study. While much of the network is in need of more studies to generate more branches, existing branches from previously performed experiments confirm the interactions shown.

Materials and Methods

Growth

Approximately 300 seedlings of each genotype were surface sterilized with bleach and ethanol, cold treated at 4°C for not less than 3 days, this was done for each biological replicate. The plant material was then grown on Al gel soak plates (Larsen et al., 1996, 2005) for 3.5 days, on both 0.0 mM and 1.5 mM Al treatments. The plates were composed of nutrient media 80 mL of 1 mM KNO_3, 0.2 mM KH_2PO_4, 2 mM $MgSO_4$, 0.25 mM (NH_4) $2SO_4$, 1 mM $Ca(NO_3)$ 2, 1 mM $CaSO_4$, 1 mM K_2SO_4, 1 mM $MnSO_4$, 5 mM H_3BO_3, 0.05 mM $CuSO_4$, 0.2 mM $ZnSO_4$, 0.02 mM $NaMoO_4$, 0.1 mM $CaCl_2$, 0.001 mM $CoCl_2$, 1% sucrose, and 0.125% Gellan gum (Alfa Azer). This provided at least 2 tubes of 100 mg of tissue, the upper limits of total RNA extraction based on recommendations of the kit. Also there would be one set of plates, one 0.0 mM plate, and one 1.5 mM plate that would be planted for a growth experiment and left to grow for 7 days. These plates would be used to confirm phenotypic change and toxicity of the media before performing RNA extraction. The tissue was collected in 1.5 mL tubes that would be immediately flash frozen by liquid nitrogen (LN_2), then stored at -80+/-10° C

RNA extraction

Total RNA extraction was performed with the Qiagen Kit RNAeasy Plant mini kit (Cat# 74904) using buffer RLT (provided). The pooled whole seedlings followed the guidelines of a max tissue weight of 100 mg per sample. The samples were removed from the -80 °C freezer and placed in liquid nitrogen. The extraction to be done via

mortar and pestle, the tools were cooled using liquid nitrogen prior to disruption. The tissue was dislodged from the tube manually, and then added to the cooled mortar and pestle to be triple ground before being added to a new 1.5 mL tube containing 450 uL of buffer RLT, which was vortexed each time ground tissue was added from the cool mortar. This is done to the keep the RNA stable while additional ground tissue is added to the buffer. Upon completion the extraction followed the protocol outlines in the Qiagen kit. Upon completion the RNA was checked for concentration and quality A260/280 approximately 2.0, and the A260/230 approximately 2.0 before proceeding to polyA capture.

PolyA capture

In this study we specifically look at those RNA that would produce protein, and to do so was done by poly(A) beads in which messenger RNA (mRNA) that contains a poly(A) tail. Using the beads from Bioo Scientific (Cat # NOVA-512979) and their protocol using the magnetic beads to capture the mRNA for processing using the RNAseq library kit. The poly(A) capture was performed using a 1.5 mL tube, rather than the plate set up, with a magnetic stand for singular 1.5 mL tubes. The poly(A) beads are first resuspended to regain uniform suspension. Next for the total RNA was calculated to ensure the proper quantity of beads was used. The amount of total RNA never exceeded the threshold of 10 ug so 20 uL was used for each sample. The kit was followed with the binding and washing, and finally to elution, to the final volume of 14 uL to be ready for RNAseq library preparation.

RNASeq library preparation

The RNAseq library was prepared using the BIOO Scientific directional RNAseq library kit for illumina sequencing (Cat # NOVA-5138-07). Using the 14 uL of mRNA the kit proceeds through using chemical fragmentation to break up the mRNA, so that the next step of first strand synthesis can occur by first annealing the generic primer to the RNA then performing a reverse transcription reaction on the sample. This leads to the creation of cDNA which can then be used to create a second strand synthesis creating DNA, which is then A-tailed and ligated to adapters to be amplified by PCR. This amplified library of mRNA fragments is then sent to UC Irvine for QC and sequencing. Each of the bead clean up steps was performed using Agent court magnetic beads (Beckman Coulter Genomics, Cat # A63880) that were stored 2-4 °C and were aliquoted out before use and allowed to rise to room temperature. A single 1.5 mL tube magnetic stand was used for all cleanup procedures.

Sequencing

Each Library was analyzed using a qubit (Thermo FIsher) and bioanalyzer (Agilent Technologies) to ensure quality and the correct size of 200-300 bp range before proceeding with multiplexing and sequencing. Sequencing was done on an Illumina HiSeq where all six libraries were multiplexed all together as one sample, and run on two lanes of the sequencer with paired end chemistry. Since for this RNAseq experiment there is more a concern with counts of the RNA rather than possible SNPs the increase data from a second lane allows us more depth as well as looking at this data as a technical replicate to help with normalization of read count data.

Analysis

The bioinformatics analysis was performed using a modified version of SystemPipeR's RNAseq pipeline [76]. Using the cut offs of a 2 fold change in gene expressions and a false discovery rate of one percent to determine DEGs. Each sequencing file is checked for quality via fastX tools [77], to determine if any quality trimming is required. The pipeline uses tophat/bowtie 2.0.14 [78], which uses a Burrows-Wheeler alignment algorithm for paired end gap aware alignments using the arguments of: only one match gene match per read, segment length of 25 bp for searching, minimum intron length of 30 bp and a max of 3 kb using 10 cores of processing power to compare the reads to the TAIR10 reference genome and the TAIR10 GFF annotation file. Using the binary alignment files (BAM files) that are generated by top hat raw counting and Reads Per Kilobase of transcript per Million mapped reads (RPKM) is performed by using R functions. This done via counts with base pair ranges of those genes, in which the length of the exions are calculated in kilobases, and the sum of the counts are are divided by one million (millions mapped). The reads per kilobase (RPK) of the exon model is then calculated with the counts divided by the millions mapped. Finally we get the RPKM via RPK divided by the kb length of the exons. While the pipeline generates a spearman plot to show the correlation on a sample wide scale.

Comparisons are done using edgeR [47] using the previously created counts and the comparisons located in the target file. EdgeR first takes the data set for the counts and does it own internal normalization, followed by statistical testing via a negative binomial distribution and biological variation correction (BVC). The comparisons used for this experiment were Col-0 treated - Col-0 untreated, *als3-1* treated - *als3-1* untreated,

and *sog1-7* treated - *sog1-7* untreated. After the list of DEGs were determined and filtered using the previously mentioned cutoffs of 2 fold change with an FDR of 1%, a venn diagram of the various samples is generated showing which genes from which samples overlap with one another. The venn diagram was created using subroutines from Girke et al, from the SystemPipeR pipeline called overlapper. which creates the table set up to use the venn diagram functions in R to produce a visual aid of the results. Next a heat map is generated using pheatmap package in R / CRAN repository, clustering is performed via the genes expression patterns of the various DEGs, while there is secondary clustering by sample.

This secondary clustering provides a good quality check and allow the analyst to gauge the quality of biological replicates where samples of the same genotype and treatment should cluster together on the heat map to have overall similarity. Using the 'hclust' function. In the case of this experiment since the chloroplast genes were not of interest, instead of generating one heat map, three were generated. The overall heat map shows both nucleic gene and chloroplast genes together, while the other to show only either nuclear and mitochondria localized or chloroplast genes. Finally the DEGs are split by sample and by change in expression. Using biomart a repository of the Ensemble database, the GO annotations are downloaded as a local database. As a note the version used in this study was the GO slim database, which only uses very general categories. A GO bar plot is created using the GO terms from the local database that map to those genes, to then aid in visualization to discover any over represented GO terms. This GO plot is broken down into three pages one for each of the main GO terms, Molecular Function, Biological Process, and cellular component. Additionally for the

manual curation, genes of interest were run through the IntAct database [68] from Ensemble to determine the interactome of the DEGs and expand the breadth of the knowledge and potential pathways. Cytoscape [69] was then to used to generate a graphical representation of the interactome.

Added modifications include command line arguments that allow the user to run the script from the command line via 'Rscript rna_seq.R targets.txt refrence_genome.fasta annotations.gff textdb' command. To optimize time and computing power, a table check is performed before starting the analysis which checks if the files created by the analysis is already created so they do not have to be recreated each time. If the file is not found, such as it has been deleted from the results directory, or not yet processed the pipeline will perform the analysis to create those files, this includes files like the txdb and the biomart file that contains the gene ontology annotations. The last modifications are specific to plants in which the heat map at the end of the analysis is divided up, the total results, those that re from the five chromosomes of *Arabidopsis* and the one that is showing just those that map to the chloroplast by using regular expressions that used 'grep' for those gene IDs that start with ATC denoting that they belong to the chloroplast "chromosome".

RPKM calculation

$$Gene\ Length = \Sigma(exon\ length)/1000$$

$$Million\ Reads\ Mapped(mrm) = \Sigma(read\ counts)/1\ x\ 10^6$$

$$rpm = counts/mrm$$

$$rpkm = rpm/Gene\ Length$$

qPCR

qPCR was performed using SYBR Green as the fluorophore, with the Bio-Rad kit iQ SYBR Green supermix (Cat #170-8882) using a 10 uL reaction with 100 ng of template and 500 nM of each primer. The standard cycling conditions of 95 °C for 3 minutes, followed by amplification of 95C for 10 seconds, and 55C for 30 seconds for 40 cycles were used for qPCR. Following this a melt curve analysis was performed beginning at 55 °C and increasing by increments of 0.5 °C every 5 seconds. The ubiquitin carrier protein (UBC) was used as the reference gene on each plate with the various genotypes that were grown the same way as those seedlings that were used for RNAseq and provide a viable comparison. Each set of reactions was done on a BioRad CFX connect machine provided at the UCR genomics core. Each plate was set up with three technical replicates, for each gene with five genotypes and corresponding treatment conditions either untreated (0.0mM Al) or treated (1.5 mM Al). Col-0 (WT), *als3-1*, *sog1-7*, *sog1-7;als3-1*, *atr-4;als3-1*.

The first three are used to validate the findings from the RNAseq, the last two are to explore the results of what would be seen with these genes for the als3-1 suppressor mutants. One lane for both the control gene and gene of interest were non template control (NTC) which allowed for control of possible primer dimers and adjusting of the threshold where actual amplification took place. While there are two ways to calculate the results of the qPCR by either comparing back the wild type genotype on the untreated condition to find the fold change in expression, there is also the calculation where each genotype on the untreated condition is used to compare to the treated conditions to see the fold change in expression of that gene for the genotype. In this

experiment the latter is used, this is due to the way the RNAseq is set up where its doing pairwise comparison each results is done with a comparison of each genotype on the treated media versus the untreated media.

The calculations follow the same algebra as the former, the only difference is the comparison. In order to control for outliers any triplicates of data points for a gene on a plate that showed a standard deviation greater than one would have the data point that most skews the data removed. After post processing the average of and standard deviation of the remaining triplicates were calculated. After which the average of the control gene is subtracted from that of the gene of interest to standardize it (dCt sample). Then for each gene the standardized average for each untreated genotype (dCt) is subtracted from the standardized averages (both treated and untreated) providing the ddCt. The standard deviation of the counts is determined from the square root of the standard deviation of the gene of interest squared added to the standard deviation of the control gene squared for the same sample condition SD dCt. Finally the fold change (FC) is calculated via two raised to the negative power of the ddCt of that sample.

$$avg\left(Ct_{unknown0.0\,mM}\right) - avg\left(Ct_{UBC\,0.0\,mM}\right) = dCt_{genotype}$$

$$avg\left(Ct_{unknown}\right) - avg\left(Ct_{UBC}\right) = dCt_{sample}$$

$$dCt_{sample} - dCt_{genotype} = ddCt$$

$$\sqrt{\left(std\left(unknown\right)^2 + std\left(UBC\right)^2\right)} = SD\,dCt$$

$$FC = 2^{-\left(ddCt\right)}$$

Primers

Gene Number	Name	Sequence
AT1G14400	UBC qPCR 5’	TCAAATGGACCGCTCTTATC
AT1G14400	UBC qPCR 3’	CACAGACTGAAGCGTCCAAG
AT1G13330	HOP2 qPCR 5’	CCAGTTTGAGATTCCAAACTCTG
AT1G13330	HOP2 qPCR 3’	CTATCGTACTCAACTTTGTCAATG
AT1G04450	RIC3 qPCR 5’	CACCACAGGAGTAGGCACG
AT1G04450	RIC3 qPCR 3’	CATGTATCGACTGATTCAACAATAGG
AT1G13980	GNOM qPCR 5’	GGTGGAGATAGCTTATGGGAGC
AT1G13980	GNOM qPCR 3’	CACTTCACAGTACTTATATG
AT1G27900	AT1G27900 qPCR 5’	CTATTCTTTATAGGGCTAG
AT1G27900	AT1G27900 qPCR 3’	GAAATTAAATGTCAGAAAATTG
AT2G36261	AT2G36261 qPCR 5’	CTCTCTCACTCACACAAAAG
AT2G36261	AT2G36261 qPCR 3’	GGATATGGATCTCAATGGATG
AT5G61570	AT5G61570 qPCR 5’	GAGATTATTAGGAGATGTCTG
AT5G61570	AT5G61570 qPCR 3’	CTGATTGTTTGAATCTTTCTC
AT5G22890	STOP2 qPCR 5’	GTGAATGCACGTGTCATTTC
AT5G22890	STOP2 qPCR 3’	CACACGAGAAGCATTGTGGGG
AT5G07620	BOBO1 qPCR 5’	CTGGCGATGATGAGTTCTATC
AT5G07620	BOBO1 qPCR 3’	CACAGACTGAAGCGTCCAAG

The Genomic Consequences of Aluminum toxicity

Introduction

Aluminum (Al) t comprises approximately eight percent of the Earth's crust by weight, making it the third most abundant element in the Earth's crust [79]. It is normally biologically inert unless the pH of the soil becomes acidic and reaches a pH of 5.5 or lower. In acidic conditions, Al speciates into its trivalent cationic form of Al^{3+}, and results in Al^{3+} being taken up by plants leading to toxicity. Previous research on this topic suggest that Al in its cationic form Al^{3+} has the ability to bind to any anionic site within the plant, the most concerning is the ability to bind to DNA. Previous research from [17] demonstrated with the COMET assay, detectable double strand breaks (DSBs) in the DNA of the plant by the amount of nucleoids with a treatment [28], treatments of Al^{3+} lead to and increase in these nucleoids; meaning that exposure to Al^{3+} has the potential to cause DNA damage.

As a global phenomenon, Al toxicity affects many parts of the world including those that are home to developing nations (Figure 1) these areas of low pH can lead to Al to speciate from its inert form to its cationic form Al^{3+} [80]. In doing so it becomes biologically available, and affects the roots as the primary tissue of plants, leading to root growth inhibition and overall lower crop yields [81] Developed and industrialized nations have access to tools, methods, and resources which these nations use to adjust the pH

of the soiland return Al^{3+} back to its inert form preventing it from being biologically available to the plant, one example of this being the addition of agricultural limestone to the soil. Developing nations often lack many of these options due to monetary costs. This monetary limitation is further compounded by the fact that many of these methods needs to be done repeatedly on the soil. When grown in the Al toxic soils, the majority of crop plants will grow smaller and produce fewer seeds, leading to an overall reduction in the yields of these crops. Thus, the inability to forestall Al toxicity leads to a much more profound problem for developing nations. Al contaminated soils can contribute to food insecurity and economic instability to regions where mitigation of this toxicity are not addressed by soil amendment or development of resistant crop varieties.

One factors that likely contributes to Al toxicity is the generation of DSBs [17]. These DSBs are hypothesized to induce DNA damage response (DDR) by the plant. There can multiple factors that contribute to DNA damage for plants including both endogenous factors, such as reactive oxygen species (ROS) [40], and environmental factors including ionizing radiation [82], heavy metals [83], and Al toxicity [84]. DNA damage kicks off a cascade of cellular and molecular responses by the plant to attempt to repair the damage. DSBs in particular can lead to genomic instability and loss of genomic information [85].

While the evidence demonstrates that exposure to Al leading to genomic damages in the form of DSBs, micronuclei, and chromosomal aberration are occurring as a result of Al^{3+} exposure [27,17], it does not provide any provide any information into any hallmarks or signatures related to the damage caused by Al exposure. This suggests two main possibilities: first there is no detectable genomic damage occurring as a result

of Al^{3+} exposure, this could be a result of any damage that is occuring being repaired in an error free and high fidelity manner. This could simply mean that a different method of analysis is required to understand to understand the consequences of these DSBs. However, if detection of Al^{3+} triggers the plants DDR in an ATAXIA TELANGIECTASIA MUTATED AND RAD3 RELATED (ATR) and SUPPRESSOR OF GAMMA1 (SOG1) dependent manner as previous research suggests [26] then these proteins are responsible to detecting DNA damage and activating the DDR. Along with the change in phenotype of Al sensitive mutants, I would interpret these findings as evidence instead some type of lasting, or heritable genetic damage, which could take many forms including but not limited to changes to nucleotides, the addition or loss of genomic information due to damage. This study seeks to characterize types and frequency of DNA damage accumulating in response to Al^{3+} treatment.

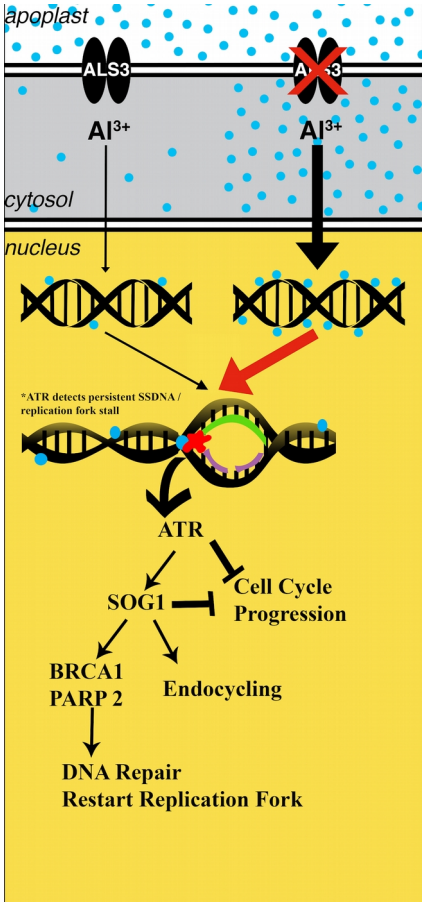


Figure 22: Current Working Model of Molecular response to Al Toxicity
The model built on the previous research in the Larsen lab [16,17,26] presents a molecular response where upon Al^{3+} enters in to the plant system it can bind to available anionic sites. When Al^{3+} is present in enough abundance it out compete the negative binding sites of the roots cell walls and make its way to Apoplast and in to the cytosol. The role of ALS3 is to pump Al^{3+} away from sensitive tissues providing protection for the root [72]. If the function of ALS3 is knocked out the amount of Al^{3+} will be significantly higher leading to the sensitive phenotype. Those Al cations have the potential to make their way to the nucleus causing DNA damage either directly or indirectly. This leads to a DDR mediated by ATR and SOG1 where the cell cycle will be halted while repairs are attempted. The plant can then undergo repair using factors such as BRCA1 and PARP2, which will be activated to repair the DNA damage and if necessary restart the replication fork. If the DNA damage can not be repair the cell may undergo endocycling rather than continuing through the normal cell cycle.

Arabidopsis as a Genetics Model

In order to study the effects of Al toxicity on the genome of plants, a well studied and curated genome is required to produce accurate results. One of the best plant models for studying Al toxicity is *Arabidopsis thaliana*. Arabidopsis serves as model organism for plants, and currently possess a well curated genome, with many knockout (KO) lines available. Arabidopsis also has a short generation time and small size which allows researchers to conduct experiments much more quickly and in a more high throughput manner than could be performed on larger crop plants.

Arabidopsis plants grows from a single fertilized seed and expands over the course of days, for the context of this experiment the plants were grown for only seven days. Below is a table of the growth of Arabidopsis based on days after planting. The seed starts as one fertilized cell and expands exponentially, first presenting a primary root and then a shoot and complex features as time progresses [86]. Many of these cells will become progenitors to other cells and different parts of the plant. Exposing early development cells to Al^{3+} could lead to changes that would be propagated throughout the entire plant. As these are diploid cells, it is likely that mutagenic or genotoxic effects from Al^{3+} will initially be caused by heterozygous mutations.

This would be dependent on what cells are affected and how early in the plant's life cycle it was affected, since due to cell division cells that undergo changes earlier in the plant's life should propagate that change more making it more detectable than change that happens later in life that could be as frequent as any other genomic change that would occur naturally.

Stage Number	Approx.number of days *	Description
0.0		Seed germination
0.1	3.0 (on plates)	Seed imbibition
0.5	4.3 (on plates)	Radicle emerges from seed coat
0.7	5.5 (on plates)	Hypocotyl and cotyledon emerge from seed coat
1		Leaf production
1.0	6.0 (on plates)	Cotyledons fully open

Figure 23: Table of Arabidopsis Growth
Data for table taken from TAIR which provides estimates of the growth of Arabidopsis over the time of the experiments from planting to 7 days [87]. This context helps provide useful estimates of the number of cells that can help provide insight into the possible genotoxic effects of Al^{3+} by using the phenotypes as these seeds .

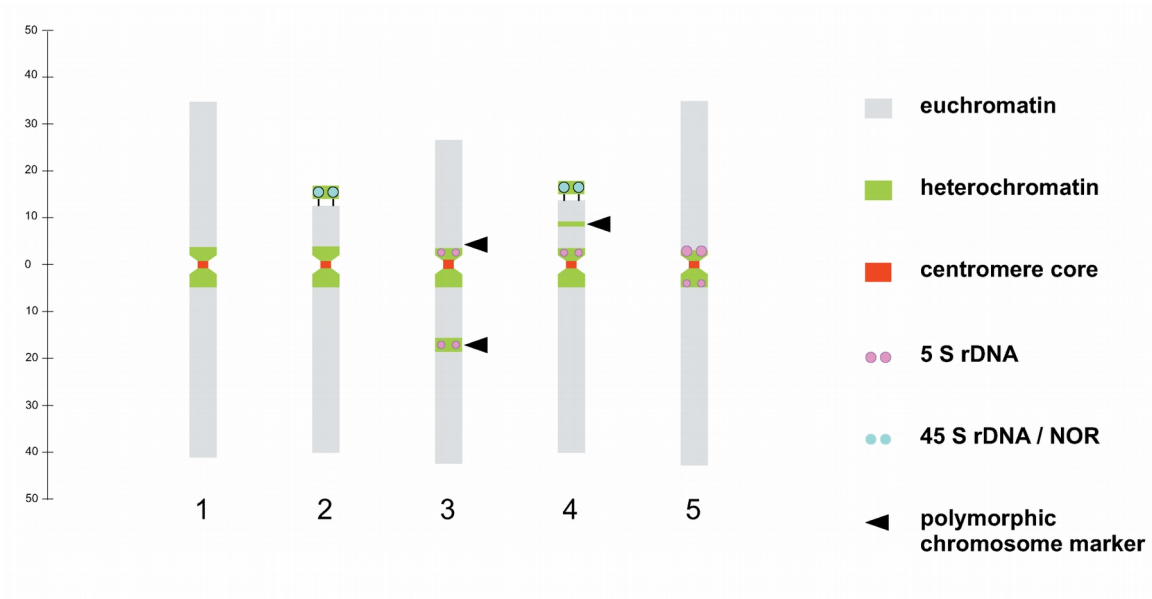


Figure 24: Arabidopsis Genome Model
Image which represents the relative size of each chromosome in micro morgans (µm) and annotated for various genomic attributes. This ideogram was produced from Arabidopsis Thaliana, Lerata ecotype. [88]

Arabidopsis has a genome size of approximately 135 Megabases (Mb) with 5 chromosomes and mitochondrial and chloroplast DNA [89]. Compared to other eukaryotes such as humans makes for a relatively small genome, this allows for greater sequencing depth and more cost effective sequencing which also allows for more biological replicates can be performed providing additional statistical confidence to the results. The sequencing depth is important since it is not known what type of damage Al^{3+} causes nor how rare this damage may be. Greater depth of sequencing allows for greater statistical confidence in the detection and identification of genomic variants arising from exposure to Al^{3+}.

Experimental Study

Previous studies provide preliminary evidence that DSBs are occuring in the presence of Al^{3+} [17] and that the plant is mounting a DNA damage response in an ATR and SOG1 dependent manner [26]. Additional studies have quantified DNA damage based on subjective methods such as the counts of micronuclei that could be visually observed on the microscope slide [27].These works have not, however, determined if the damage being caused by treatment with Al^{3+} can be quantified and if there any signatures related to the damage.This study seeks to answer that question and provide quantitative evidence that Al^{3+} leads to DNA damage that increases in a dose dependent manner that can be detected using Next Generation Sequencing (NGS) and can be subsequently quantified. I hypothesize if I treat these plants with Al there will be DNA damage, if so then this damage should be occuring in a dose dependent manner, detectable using NGS by comparing samples of different treatments. However if my hypothesis is incorrect, the results will lead to the acceptance of the null hypothesis that there is no quantifiable genomic damage occurring or at the very least lead to a revised hypothesis requiring another method of detection.

To ensure that our lines have not been previously exposed to Al^{3+}, two virgin seed lines (seeds that have never been exposed to Al) were obtained from the Arabidopsis Resource Center (ABRC) at Ohio State. One of Columbia wild type line (P14), and one the *ALUMINIUM SENSITIVE 3* (*ALS3*) mutants, *als3-3* which is KO line that possesses a segment of transfer DNA (t-DNA) in an exon of *ALS3* knocking it out its function as an ABC-type transporter, which helps to regulate ions in the plants cells [72].

Knocking out this gene leads to a hypersensitive phenotype in response to exposure from Aluminium which is characterized by terminal differentiation of the root tip and overall smaller plant growth when exposed to Al^{3+}. This *als3* mutant is equally or more sensitive than *als3-1* [72]. The *als3-1* allele contains a point mutation in *ALS3* resulting in a change of amino acids from a serine to a lucine knocking out its function [90]. Due to its hypersensitivity to Al^{3+}, *als3-3* should generate much more clear results if there is detectable DNA damage, while P14, which is not as sensitive, could have only slight differences between treated and untreated plants leading to unclear, even confounding results. The predicted stark results of untreated vs treated of the sensitive mutant will in turn help identify what patterns or types of damage, if any, can be detected, and use it to guide the analysis of the wild type results.

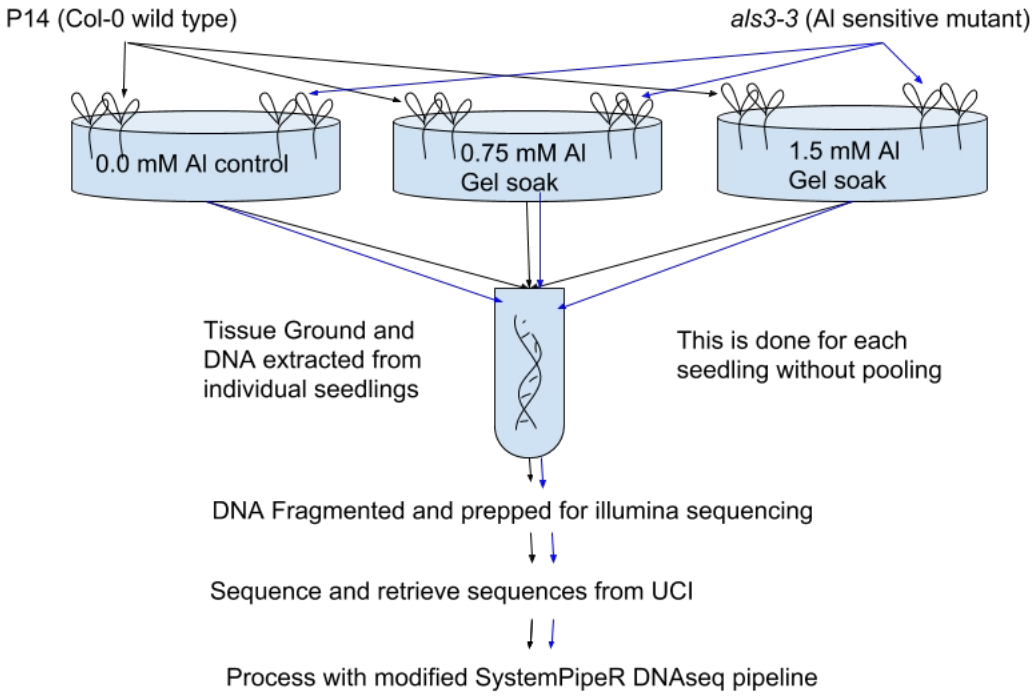


Figure 25: Illustration of Experimental Flow
This illustration visualizes the flow of the experiment from planting of the seeds for both lines (P14 and *als3-3*), and flowing down to the DNA extraction, the samples are sent UC Irvine genomics core for fragmentation by mechanical shearing. After the samples are returned they are processed using a NEB Ultra library prep kit for Illumina sequencing. After the libraries are sent to UC Irvine Genomics core, and the quality checks for size of fragments and concentration of the library are performed using a bioanalyzer. After which the libraries are multiplexed and sequenced. The fastq raw sequencing files are then retrieved, and analyzed using a modified version of SystemPipeR .

To perform the experiment, gel soak media was generated with 0.0 mM, 0.75 mM, and 1.5 mM Al soak solutions all of which are at pH 4.2. Each plate contained 10 surface sterilized seeds of each genotype, meaning 10 seeds of P14 and 10 seeds of *als3-3* each were planted on each plate soaked with 0.0 mM, 0.75 mM or 1.5 mM Al. The plants were allowed to germinate on the media and grow for 7 days, after which whole seedlings were collected. Whole seedlings were used for consistency as the *als3-*

3 seedlings size and amount of DNA for extraction was very small, some less than 200 ng total DNA, under from plants under the treated conditions.

Recovering so little DNA lead to difficulties keeping libraries consistent between replicates, due to *als3-3* having undergone terminal differentiation of the root by the time the tissue was collected so that one sample may have more total DNA than another. However, each biological replicate was processed identically. The seedlings were grown up and total plants were macerated for DNA extraction for Illumina sequencing library preparation. The DNA samples were fragmented via mechanical shearing using a Covaris S2 sonicator to shear the DNA with a target of 300-400 base pair (bp) fragments. Illumina sequencing libraries were prepared from these sheared DNA samples using SPRI beads and NEB ultra DNA library kit (details in materials and methods). Libraries were multiplexed at three libraries per flow cell lane on a HiSeq 2500 to obtain a target of 25 Million 2x100 paired reads per library and a target genome coverage of ~90X coverage.

Empirical evaluation of the sequence reads demonstrated that libraries generated an average of 89X coverage across the genome. Analysis using a optimized SystemPipeR [76] DNAseq pipeline, with additional code to decrease the required time of the analysis. Additional functions for more detailed analysis were also added to produce the basics of the results following the workflow in Figure 25. These intermediate results were processed further to remove variants that were due to differences in the reference genome sequence and identify variants most likely to have been caused by exposure of Al^{3+}. The identified variants were processed to test for patterns that related to factors such as: the overall rate of changes, type of mutation, or genomic location of variations

that were dependent on Al^{3+} exposure concentration. Due to its hypersensitivity, it was expected there would be an excess of mutations in *als3-3*. Due to this hypersensitivity to Al^{3+} creating such an extreme response, it also means that the results of the two genotypes are unable to be compared. These experiments and analyses allow a framework to test for genomic consequences of Al^{3+} sensitivity and exposure. Based upon the previous research conducted on this topic, I expect to conclude that there is genomic damage occurring from Al^{3+} that can be detected using NGS, and that there will be an increase in damage correlating with an increase in the concentration of the treatment. Additionally I hope to find a pattern of some kind relating to the damage (a genomic signature) to allow for the basis of future research to continue on this topic as to the potential reason why this damage occurs, and how the plant attempt to repair it.

Results

Sequence Analysis Pipeline

The pipeline developed an applied in this study, which builds on the backbone of SystemPipeR [76], allows for more reliable, and reproducible results.This due to the fact that the pipeline already has programs included that have already been tested with appropriate parameters determined, and published as opposed to trying to determine these settings through trial and error. The use of standardized running parameters within the pipeline integrating existing third party software suites creates better reproducibility between samples and replicates especially for the core of the genomics analysis: read mapping, calling variants, and quality filtering. Having the pipeline as the backbone of

the analysis allowed the development of subroutines that were focused more specifically for the downstream analytics and reports. To ensure that these custom scripts were working correctly synthetic sequences were generated, in doing so each function could be tested for its validity, and correct results prior to use with the genomics data. The custom subroutines were tested using a known subset of data to ensure their accuracy prior to being used in the analysis and were tuned there after using the large data sets of both P14 and *asl3-3*. The additional modules and steps added to my custom DNA-seq pipeline generated graphs and figures, controlled the curation and added improved functionality, such checking for files that have already been created with the ability to skip time consuming steps if already completed, thus decreasing the overall time for analysis. While tailored to *A. thaliana*, these additional subroutines could be used for other organisms with different number of chromosomes and other genomic features such as not having chloroplast.

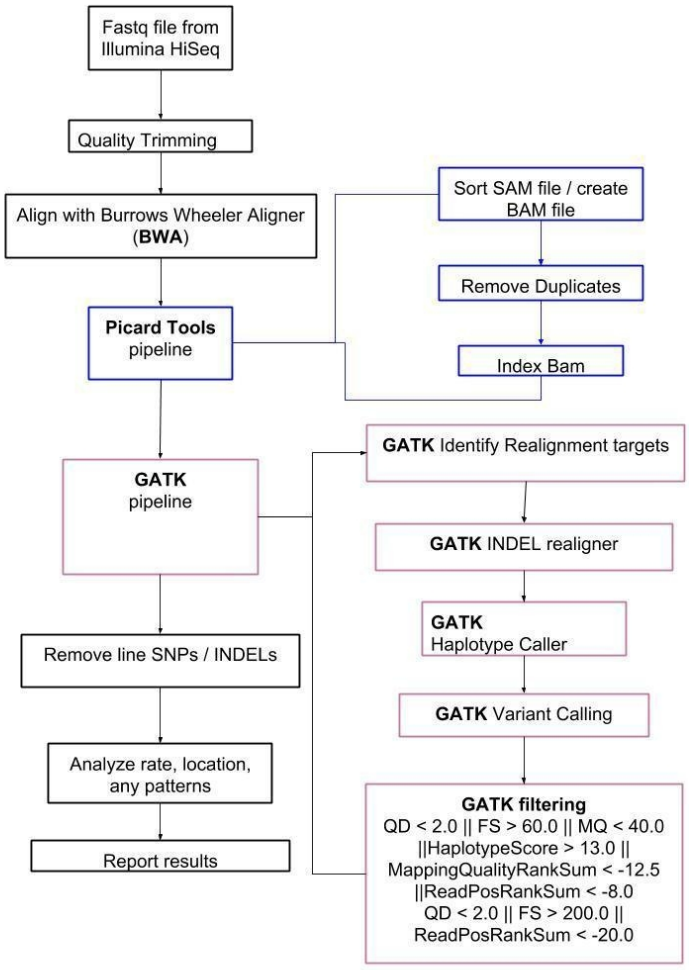


Figure 26: Genomic Pipeline Flow Diagram

The raw reads are trimmed for quality using Trimmomatic [91] followed by alignment to the reference genome using Burrows-Wheeler Aligner [92]. Duplicate reads are marked and moved using Picard Tools [93] and stored as binary alignment (BAM) file. Regions around identified Insertion/Deletions (INDELs) are realigned following Genome Analysis ToolKit (GATK) best practices [94]. Initial variant calling was performed with HaplotypeCaller followed by filtering using the cutoffs listed above following GATK best practices implemented inr following the established GATK pipeline as the underlining part of SystemPipeR's DNAseq analysis. The variants that passed the quality control methods in the VariantAnnotation package were the resulting output for candidate mutations [95]. The General Features Format (GFF) file which encodes the gene location annotation of A. thaliana obtained from TAIR (TAIR10) was used to identify where the genomics landed in term of the gene models, and the potential impact they could have on the organism. This classification/analysis provides context to the variants likely caused by exposure to Al^{3+} and potential impact to genome integrity of the plant.

Filtering of Genomic Variants

To identify those variants that are caused by exposure to Al and reduce false positives, a binary table was generated from the high quality variants filtered by GATK (Table 2). Each set of detected variants from the various concentrations: Control (0.0 mM) low treatment (0.75 mM) and high treatment (1.5 mM). Variants with the identical genomic location, and change were investigated to determine a rate of change for the different treatments of Al^{3+}. The goal of the study is to determine if there is a dose dependent increase in mutations with the treatment, this table facilitated the identification and filtering of variants associated with exposure to Al^{3+}. Additionally it allows for greater focus on those Al associated variants that pass filters designed to remove low quality variants and provides context of the sample population to help guide further analysis.The goal of this table to provide clarity of which variants are important and represent real changes related directly to the treatment and to screen out those that are not to ensure that the results that follow are based the results of the treatment by Al^{3+}.

0.0 mM Al^{3+}	0.75 mM Al^{3+}	1.5 mM Al^{3+}	Action / Result
1	0	0	Basal Rate of mutation
1	1	0	Intermediate group
1	0	1	Intermediate group
0	1	0	Al - change to investigate
0	0	1	Al - change to investigate
0	1	1	HFC
1	1	1	line change / difference from TAIR

Table 2: Binary Variant Assignment Table
The columns of this table correspond to the treatment of samples. A zero indicates a variant is not present, while a one indicates that is present. This binary method of grouping the variants allows us to focus on just those variants of relevance. The control media soaked in no Al (0.0 mM Al^{3+}), Those samples that were in the low treatment (0.75 mM Al^{3+}) condition in which is presented in the second column, and lastly the high treatment (1.5 mM Al^{3+}) is for the samples that correspond to the samples coming from the media soaked in 1.5 mM Al. To determine the basal rate of variation as well as which variants should be processed further due to relation to Al^{3+} the table shows the following; it combines all filtered variants from each treatment (SNP and INDELs were handled independently).

These samples were compared to the reference genome using read mapping, inherent in this method, the sequences will contain some differences which will be inherent variants (line variants). These line variants can be related to the time since the TAIR10 reference genome was sequenced, additionally these changes could represent sequencing errors amount all the samples, both of which could be expected to be present in the analysis. The *A. thaliana* accession lines used in the experiment could have accumulated lineage-specific mutations, including a heterozygous allele becoming homozygous for a base that differs from what is recorded for that position in the reference genome. This basal rate of change emanating from changes unique to only

untreated samples when compared back to reference genome, likely resulting from difference in genotype or population changes provides us with a baseline for comparison (See following tables Table 3 and Table 4 for the Binary Rates). My analyses found 18 percent difference between the samples and the reference genome as the baseline for both SNPs and INDELs, as such, values above this threshold indicate that those changes could be due to Al^{3+} exposure (signifying Al^{3+} could the cause of this increase). This 18 percent includes many heterozygous variants, and was intentionally sensitive since there no known signature of damage generated by Al^{3+}. Next we see the line intermediates ranging from seven percent in the indels to 12 percent in the SNPs. The high frequency changes (HFC) range from three to 6 percent representing changes found in both level of treatment but not in the control appearing as a rare event compared to the rest of the categories. The line changes which were common changes among all three samples accounted for 25 to 30 percent of the variants. Since aforementioned categories can not be directly attributed to exposure of Al^{3+}, the remainder of the study focuses on those variants that were categorized in to to the remaining types of changes that corresponded to appearing in just those types of treatment. This comprised 16 to 17 percent of the variants in the SNPs for the low (0.75 mM) treatment and 19 to 20 percent in the high (1.5 mM treatment). As for the INDELs the low treatment accounts for 15 to 19 percent of the changes and the high treatment is represented by 21 to 22 percent of the changes.

This was done using the binary table which was generated through R code, created as a custom subroutine using the preceding table for the logic behind the table. These are the unique counts of each variant as they fall onto the table (Table 3, Table

4). Variants that were observed more than once in the same type of sample were called only once. This limiting to unique sites rather than mutations is important in calculation of rates which will differ from the total observed mutation counts when variants are analyzed independently.

	P14 SNP frequency		*als3-3* SNP frequency	
Group	Total	Percent	Total	Percent
Basal Rate	5619	18.27 %	6074	18.22%
Line intermediate	3881	12.62%	4019	12.06%
Al changes 0.75	5154	16.76%	5905	17.71%
Al changes 1.5	5832	18.96%	6922	20.76%
HFC	1761	5.73%	1967	5.90%
Line Changes	8507	27.66%	8449	25.34%
Totals	30754	100%	33336	100%

Table 3: Binary Rates for SNPs in Each Genotype
Using the categories outlines in Table 2, each genotype is independently categorized to the various variant categories, and their frequency counted. This table ignores duplicates, and simply determines presence of, or absence of, different types of variants. For each genotype, the categories are displayed showing the total number of unique binary changes, followed by the percent of the total.

	P14 INDEL Frequency		*als3-3* INDEL Frequency	
Group	Total	Percent	Total	Percent
Basal Rate	1900	18.58 %	2159	18.44%
Line intermediate	800	7.82 %	821	7.012%
Al changes 0.75	1634	15.98%	2295	19.60%
Al changes 1.5	2205	21.57%	2652	22.65%
HFC	356	3.48%	511	4.36%
Line Changes	3329	32.56%	3270	27.93%
Totals	10224	100%	11708	100%

Table 4: Binary INDEL Rates for Each Genotype
Using the categories outlined in Table 2, each genotype is independently categorized to the various variant categories, and their frequency counted. This table ignores duplicates, and simply determines presence of absence of different type of variants. For each genotype, the categories are displayed showing the total number of unique binary changes, followed by the percent of the total.

Interestingly, from the results show evidence that treatment with the 0.75 mM gel soak plates, actually decrease the rate of both SNPs and INDELs in these plants (Tables 3 and 4). In contrast, growing samples on the 1.5 mM gel soaked plates, equivalent to about 100 uM of Al^{3+} shows a consistent increase above the baseline, though very small in the case of the P14 where there were only 0.7 percent increase in SNPs compared to the 2-4 percent increase over the baseline of the other samples. Based on the raw counts, the magnitude of Al induced changes in the 1.5 P14 and Al changes 0.75 of *als3-3* seem to indicate that these genotypes cannot be compared.

This due to the *als3-3* hypersensitivity where the response to the low dose mimics that of high dose in P14, demonstrating just how different these two lines are under the pressure of the treatment. This is also seen in the morphology of the plants where P14 treated with a high dose will undergo terminal differentiation but *als3-3* will undergo these changes at only 0.75 mM.

The line intermediates were examined further to determine the likelihood that these variants were real changes and a result of Al exposure. For instance, assessing the possibility a line variation that was heterozygous in the samples of used for all three levels of treatment, could then have a change occur in one type of sample resulting in a variant that now matches the nucleotide at that position in reference genome. In doing so it would no longer be called a variant and only exist in the untreated and one type of treated sample. Using these variants, the goal is to provide clarity regarding the molecular consequences they could represent for the plant, and scrutinize them for any potential information they could provide about the results Al^{3+} exposure. Shown in Table 5 the majority of these variants are heterozygous representing 91 to 99 percent of the total variants in each line and variant type (SNPs or INDELs), which suggests that these variants that were categorized as line intermediates are more likely in a heterozygous state. This could mean that these variants were not present in all the samples to meet the criteria of a line variant as outlined in the Binary Variant Assignment Table (Table 2).

As previously stated the goal of inquiry is to determine if these variants in Table 5 could provide any useful insights into the results of Al exposure. Based on these results it appears they do not as its more likely that these changes were simply the result of population variation, not do to any response post exposure to Al^{3+}.

Allele Frequency of Al - Intermediate Variants			
Sample	Type	Counts	Percent
P14 SNPs	Heterozygous	9773	99.13%
P14 SNPs	Homozygous	86	0.87%
P14 INDELs	Heterozygous	1813	91.84%
P14 INDELs	Homozygous	161	8.16%
als3-3 SNPs	Heterozygous	7130	98.90%
als3-3 SNPs	Homozygous	79	1.10%
als3-3 INDELs	Heterozygous	1872	93.69%
als3-3 INDELs	Homozygous	126	6.31%

Table 5: Al - intermediate variants breakdown
This table shows the breakdowns of the Al - intermediate samples, the percentage of these samples was higher than originally expected. As the table depicts each variant type and genotype combination was examined for what state either homozygous or heterozygous the variant determined to be present. Of each pair of genotype and variant combination the percent of the total of those variants was determined as to how much it is heterozygous vs. homozygous.

To Identify variants linked to exposure from Al^{3+} with confidence from each genotype, the variants categorized as Al-dependent changes at both the 0.75 mM and 1.5 mM were tested for statistical significance. I applied a Fisher 2x3 table to compare genotype (*als3-3*, P14) and treatment (Control, 0.75mM 1.5mM) Chi-Squared Test to assess significance since number of observations exceeded the utility of Fisher's exact test. This test provides insight showing that there is a significant difference for these variants in the treated conditions compared to the control conditions, for both the SNPs (Table 6) and the INDELs (Table 7).

	C_1 (0.0 mM)	C_2 (0.75 mM)	C_3 (1.5 mM)	Totals
R_1 (*als3-3*)	6074	5905	6922	18901
R_2 (P14)	5619	5154	5832	16605
Totals	11693	11059	12754	35506
Chi-Square	13.44		Df = 2	P = 0.0012007

Table 6: SNPS 2x3 table
This two by three conditional table from VassarStats website [96] compares the frequency of the SNPs for both genotypes that were assigned to the basal rate of change, the Al-change 0.75 mM and Al-change 1.5 mM categories shown in table 2. The basal rate is the first column followed by the Al variant determined to be from the 0.75 mM samples, and the column from the 1.5 mM samples. The first row is the *als3-3* samples and the second is the P14 samples. In this case using the Chi-Square test gives a $p < 0.01$ showing that there is significance between the doses among the samples.

	C_1 (0.0 mM)	C_2 (0.75 mM)	C_3 (1.5 mM)	Totals
R_1 (*als3-3*)	2159	2295	2652	7106
R_2 (P14)	1900	1634	2205	5739
Totals	4059	3929	4857	12845
Chi-Square	23.66		Df = 2	P = 0.000007

Table 7: INDEL 2x3 Table
This two by three conditional table from VassarStats website [96] compares the frequency of the INDELs for both genotypes that were assigned to the basal rate of change, the Al-change 0.75 mM and Al-change 1.5 mM categories shown in table 2. The basal rate is the first column followed by the Al variant determined to be from the 0.75 mM samples, and the column from the 1.5 mM samples. The first row is the *als3-3* samples and the second is the P14 samples. In this case using the Chi-Square test gives a $p < 0.01$ showing that there is significance between the doses among the samples.

A 2x3 table uses a odds ratio to examine the inferred changes caused by the Al treatment tests for significant changes that can be attributed to the treatment or the genotype. These results suggest that there is significant change, that comes from the mutant *als3-3*, but also the 1.5 mM treatment. The 1.5 mM treatment shows an increase in both INDELs and SNPs, in terms of genotype however als3-3 shows a consistent increase in INDELs from the basal rate up to the high treatment levels. These findings provide evidence of the significant changes and require further analysis to determine what about these treatment and genotypes the Al^{3+} might be doing to create such a significant response.

SNPs

The SNPs were analyzed as to whether any patterns in type of change, location in the genome, or genomic context (Exon, Intron, or Intergenic). The total numbers and per-feature summary of observed SNPs across treatment conditions show no significant changes when tested via Analysis of Variance test (ANOVA).

	Degrees of Freedom	Sum of Squares	Mean Squared	F value	p value
SNP type	11	1115460	101405	19.231	<2e-16 *
Al treatment	1	4845	4845	0.919	0.341
Type : Treatment Interaction	11	4297	391	0.074	1.000
Residuals	72	379648	5273		

Table 8: P14 Total ANOVA Results
An ANOVA was performed for the P14 samples to determine if further investigation was required to look for dose dependent changes (shown by the Al category). The various rows represent different factors in this analysis such as: “SNP type” which is the category for the different categories of SNPs detected by the analysis for every permutation of changes for each nucleotide. The Al treatment is merely if the samples were treated with high or low Al^{3+}. Lastly the interaction looks to see if there is any interaction between the other two factors. Based on the findings while there are significant differences within the type of genomic variants, they can cannot be determined to be caused by exposure to Al^{3+}. The significance is shown with * to denote $p < 0.05$.

	Degrees of Freedom	Sum of Squares	Mean Squared	F value	p value
SNP type	11	1511139	137376	27.841	<2e-16 *
Al treatment	1	15965	15965	3.236	0.0762
Type : Treatment Interaction	11	22407	2037	0.413	0.9459
Residuals	72	355266	4934		

Table 9: *als3-3* total ANOVA Results
An ANOVA was performed for the *als3-3* samples to determine if further investigation was required to look for dose dependent changes (shown by the Al treatment category). The various rows represent different factors in this analysis such as: “SNP type” which is the category for the different categories of SNPs detected by the analysis for every permutation of changes for each nucleotide. The Al treatment is merely if the samples were treated with high or low Al^{3+}. Lastly the interaction looks to see if there is any interaction between the other two factors. Based on the findings while there are significant differences within the SNP type of genomic variants, they can cannot be determined to be caused by exposure to Al^{3+}. In the case of *als3-3* with the hypersensitive phenotype the dosage of Al^{3+} produces stronger results, however they are still not significant even at $p < 0.5$. The significance is shown with * to denote $p < 0.05$.

Allele frequency

In analyzing the allele frequency of Al-specific variants, the majority of the changes were heterozygous. Due to the growth conditions it is expected that many of the mutations that will occur from the Al treatment will occur after germination, and will very rarely affect progenitor cells, leading to the majority of the changes being somatic in nature. Which means that the majority of the mutations that are likely to occur as a result of Al^{3+} treatment will be determined as heterozygous since the whole seedling is being sequenced. In the previous section the data showed, overall, that the changes were predominantly heterozygous changes. To investigate if this trend was reliable, the data from the SNPs were examined as a subset of the total data set.

P14 Al Induced SNP Allele Frequency				
Allele Freq	0.75 mM counts	0.75 mM %	1.5 mM counts	1.5 mM %
Homozygous	26	0.46	46	0.73
Heterozygous	5603	99.54	6264	99.27
Total	5629	100	6310	100

Table 10: P14 Al Induced SNP Allele Frequency
The breakdown of the total counts of variants based on the allele frequency associated with that change based on the raw counts of those variants that grouped into the Al induced change category. Containing both the 0.75 mM and 1.5 mM treatments for comparison along with the corresponding percentage.

als3-3 Al Induced SNP Allele Frequency				
Allele Freq	0.75 mM counts	0.75 mM %	1.5 mM counts	1.5 mM %
Homozygous	38	0.60	68	0.90
Heterozygous	6250	99.40	7459	99.10
Total	6288	100	7527	100

Table 11: *als3-3* Al Induced SNP Allele Frequency
The break down of the total counts of variants based on the allele frequency associated with that change based on the raw counts of those variants that grouped into the Al induced change category. Containing both the 0.75 mM and 1.5 mM treatments for comparison along with the corresponding percentage.

Based on these results, less than one percent of SNPs detected are homozygous in both treatments. The remaining 99% of the SNPs are heterozygous is consistent with the assumption that most of these changes are occurring in somatic tissue and not in progenitor cells if they are induced by Al. They also show an overall increase in a dose wise manner, lending support to the hypothesis that Al^{3+} could be causing detectable DNA damage in a dose dependent manner. There was also an interesting correlation that the 1.5 mM treated samples of the P14 line seems similar in magnitude as the *als3-3* 0.75 mM samples. This could be due to the hypersensitivity of the *als3-3* line or due to the overall genomic changes that are present in *als3-3* samples due to the mutagenesis that led to the generation on this mutant line.

Genomic Hotspots

To gain a broader perspective of these changes, the genome was divided set into discrete 100 kilobase (kb) units known as ‘bins’. This binning procedure includes a final bin that covered any genomic region that was left over after the rest of the chromosome was divided up. The following plots illustrate the raw counts of those

variants based on the bins (Figures 26, 27). Due to the size of the chromosomes and rare occurrences of Al variants, the binning results was represented as a chromosomal heatmap comparing low (0.75 mM) and high (1.5 mM) treatment, where blue is the lowest going toward yellow at 50% and high shown in red (the max value for the chromosomal data). This is done specifically within the chromosomes to look regions where there is an increase in variants for that bin, and identify possible regions of interest for future studies.

The areas of the heat maps for each genotype show some interesting pieces of information. As I hypothesized, *als3-3* is more sensitive, and demonstrated by the increased frequency and intensity of the changes in across chromosomes leading to the identification of genomic hotspots. These hotspots include regions of the genome where multiple genomic variations are occurring and accumulating determined by chromosomal locations where there was in increase in the frequency of variants in a bin, (blue to yellow, or yellow to red). While this is a useful method of visualizing the impacts of the variants on various locations throughout the genome, it does lack any testable results to determine if the results are significant. However, this does provide a starting point for future testing of genomic locations where consistent genomic hotspots are seen to potentially validate these findings.

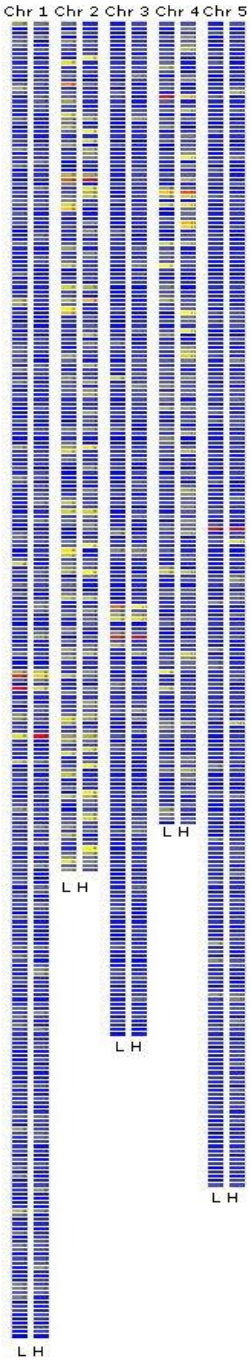


Figure 27: P14 Genomic Plot of SNPs in 100 kb Bins Starting from the top, the heat map shows chromosomes 1-5 and the average rate of detected Al SNPs across each chromosome. Those locations ranging from minimum of zero in the dark blue to the max values in red. Each chromosome has two columns the left being the results of the 0.75 mM experiment (low treatment), and the right column the results of 1.5 mM experiment (high treatment).

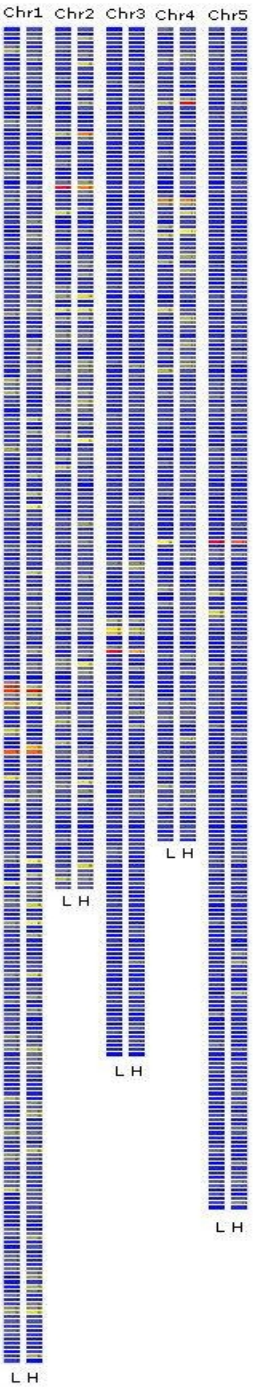


Figure 28: *als3-3* Genomic Plot of SNPs in 100 kbp Bins
Starting from the top, the heat map shows chromosomes 1-5 and the average rate of detected Aluminum SNPs across each chromosome. Those locations ranging from minimum of zero in the dark blue to the max values in red. Each chromosome has two columns the left being the results of the 0.75 mM experiment (low treatment), and the right column is the results of 1.5 mM experiment (high treatment)

Variant Breakdown

A primary hypothesis predicts that *if Al is causing changes then there should be some form of dose dependent response*. If there is a dose-dependent response I would expect that there would be an increase from the rate of change in the low treatment to that of the high treatment, from those SNPs for both P14 and *als3-3* that were found to be associated with Al exposure. To make the comparison more useful the counts of the changes were normalized using the chromosomal length, allowing for comparison across the genome as well as between treatments. Overall in comparison between the two treatments the rate of change actually decreases with an increase in the treatment.

Chr	Length	0.75 mM Changes	0.75 mM Rate of Change (SNPs / 10kb)	1.5 mM changes	1.5 mM Rate of Change (SNPs / 10kb)
1	30427671	286.33	0.19202335	272.67	0.1568652057
2	19698289	205.67	0.212087578	180	0.1721604995
3	23459830	215.33	0.1913802342	218.67	0.1700188721
4	18585056	175.67	0.1950318877	163.33	0.1462366011
5	26975502	259	0.1960130937	240.67	0.1442009585

Table 12: P14 Overall SNP Changes
This table shows a breakdown based on the concentration of those SNPs that were associated with the treatment to aluminum as an average count and the rate of change to check for any increases rate of change. The length of each chromosome is provided for reference.

Chr	Length	0.75 mM Changes	0.75 mM Rate of Change (SNPs / 10kb)	1.5 mM changes	1.5 mM Rate of Change (SNPs / 10kb)
1	30427671	355	0.2561497281	266.67	0.1794333495
2	19698289	228.33	0.2517264237	172.33	0.1790745428
3	23459830	277.33	0.2585850228	222	0.1953684649
4	18585056	216.33	0.2518299644	183.33	0.2035430052
5	26975502	319	0.2579136505	240	0.1860545884

Table 13: *als3-3* Overall SNP Changes
This table shows a breakdown based on the concentration of those SNPs that were associated with the treatment to aluminum as an average count and the rate of change to check for any increases rate of change. The length of each chromosome is provided for reference.

Interestingly the 0.75 mM treatment is very consistent with the rate of changes of the SNPs, for example the P14 samples all the SNPs in the low treatment are about 0.2, and in *als3-3* the are all about 0.25. Where as the high treatment in both genotypes fluctuates much more ranging from 0.14 - 0.17 in P14 and 0.17 - 0.20 in *als3-3*. Overall, though the rates of change fluctuate between low and high dose, there is no indication of any dose dependent increases that could be associated with the increase in dosage of the Al^{3+} with in each genotype for the SNPs.

Rates of Change

If the exposure to Al^{3+} results in the plant mounting a DNA damage response, Al^{3+} could be leading to a change in bases from one base to another in a dose dependent manner. While both are shown in raw counts, the overall order of magnitude can be compared between the doses of Al^{3+} to help look for overall trends.

	P14 0.75 mM SNPs				P14 1.5 mM SNPs			
Base	A	C	G	T	A	C	G	T
A	0	19	30.5	216.25	0	23.5	42	220.25
C	242	0	10	212.5	285.75	0	12.25	232.25
G	176.25	10.75	0	298.25	202.75	10	0	308.25
T	147	29.25	15.5	0	181.25	39	20.25	0
Base	A	C	G	T	A	C	G	T
A	0.00%	7.15%	11.48%	81.37%	0.00%	8.22%	14.70%	77.08%
C	52.10%	0.00%	2.15%	45.75%	53.89%	0.00%	2.31%	43.80%
G	36.32%	2.22%	0.00%	61.46%	38.92%	1.92%	0.00%	59.17%
T	76.66%	15.25%	8.08%	0.00%	75.36%	16.22%	8.42%	0.00%

Table 14: P14 Rate of Nucleotide Changes
Rate of change (frequency of variant detection per sample of that treatment) between treatments with the reference base on the left column and the number of changes to the alternate based list on the top of each column. The low treatment of 0.75 mM is shown in blue and the high treatment of 1.5 mM is show in yellow. The second half of the table shows row normalized values as percent of total variants for the reference base for each treatment.

	als3-3 0.75 mM SNPs				*asl3-3* 1.5 mM SNPs			
Base	A	C	G	T	A	C	G	T
A	0	23	38.25	234.25	0	24.25	39.25	286.5
C	271.25	0	14	236.5	332.25	0	12.75	279
G	184	15.25	0	314.5	252.25	12	0	392
T	180.75	39.25	21	0	196	35.25	20.25	0
Base	A	C	G	T	A	C	G	T
A	0.00%	7.78%	12.94%	79.27%	0.00%	6.93%	11.21%	81.86%
C	51.99%	0.00%	2.68%	45.33%	53.25%	0.00%	2.04%	44.71%
G	35.82%	2.97%	0.00%	61.22%	38.44%	1.83%	0.00%	59.73%
T	75.00%	16.29%	8.71%	0.00%	77.93%	14.02%	8.05%	0.00%

Table 15: *als3-3* Rate of Nucleotide Changes
Rate of change (frequency of variant detection per sample of that treatment) between treatments with the reference base on the left column and the number of changes to the alternate based list on the top of each column. The low treatment of 0.75 mM is shown in blue and the high treatment of 1.5 mM is show in yellow.The second half of the table shows row normalized values as percent of total variants for the reference base for each treatment.

In comparing the rates of nucleotide changes between 0.75mM and 1.5mM treatments for each genotype, each variant type which is determined from what reference genome says the base should be (shown in white of tables 14 and 15) compared to the changed base in the shown in blue for the 0.75 mM treatment, and yellow for the 1.5 mM treatment. In looking at the overall magnitude of the of the changes, comparing the blue to the corresponding yellow boxes the overall magnitude seems to be constant with a small exception. The rates of changes of bases to either A or T seem to increase with increase with Al^{3+} dosage. In comparison the changes to either C or G from the reference does not show consistent increases or correlation with the dosage of Al^{3+}. However these are merely trends as there was no significant bias in change based on the ANOVA.

The P14 line appears to show dose-dependent increase in G do show an increase. However this could be independent of Al since this isn't seen in *als3-3* as well. Overall, these results provide the beginning of a pattern that could suggest that Al^{3+} associated variants are more likely to end up being A or T bases vs other nucleotides.

Transitions and Transversions

Transitions and transversion were examined as part of the analysis in to the changes in nucleotides as a possible result of Al^{3+} exposure. This relates directly with the rate of changes in the SNPs and instead of looking at each nucleotide independently instead looks at changes based on the nucleotide groups, purines (A or G) and pyrimidine (C or T). These rates of change were examined for purine to purine, or pyrimidine to pyrimidine (Transitions) versus the rate of change of purine – pyrimidine or

pyrimidine to purine (Transversion). These changes are measured as a possible sign of mutagenicity or repair bias as a result of Al^{3+} exposure. This table (Table 16) shows that in wild type we can see the ratio increasing in a dose dependent manner with the concentration of Al^{3+}. Statistically without any prior data to say otherwise there should be an equal probability should be about 0.50 of any change leading to either a transition or transversion. However based on other studies done with using TAIR 8 reference genome, the results show a ratio of Transitions / Transversions to be 2.4 [97], in terms of the results of table 16 this would be converted to 0.41. Even though the raw counts do show a dose dependent increase with concentration of Al^{3+} there was almost an equal distribution of changes to A or T, the only bias in the changes of C and G to A and T, is there seems to be preference for G to become a T. A prefers to become a T and vice versa, however C would be come either A or T.

	0.75 mM P14	1.5 mM P14	0.75 mM *als3-3*	1.5 mM *als3-3*
Transitions	451.75	522.25	501.75	610.5
Transversions	962	1066.75	1079.75	1288.25
Ts_Tv_ratio	0.482757	0.545814	0.50029025	0.49051875

Table 16: Transitions and transversions between treatments
The raw counts here are presented as an average of the counts among each type of sample (genotype, treatment concentration) combination.

Predicted Impact

As part of the analysis, the program SNPeff [98] was used to predict the possible transcriptional outcomes as a result of the exposure to Al^{3+}. This will provide a prediction of whether Al^{3+} could lead to functional mutations, changes that are in coding regions in the plant. If Al^{3+} could produce functional mutations these predictions could provide a better understanding of the possible genomic consequences as a result of Al^{3+} exposure. Additionally, the results are again broken down by exposure to allow for the identification of a dose dependent changes to the organism. Overall, the results of this analysis for the SNPs seem to be consistent with the categories showing the highest percent of predicted changes being regions of least effect, such as intergenic regions of upstream or downstream gene variants.

	P14 0.75 mM		P14 1.5 mM		*als3-3* 0.75 mM		*als3-3* 1.5mM	
Type	Count	Percent	Count	Percent	Count	Percent	count	percent
DOWNSTREAM	1,943.67	36.13%	2,290.33	35.57%	2,620.00	37.03%	3,035.75	35.43%
EXON	109.67	2.04%	134.33	2.09%	128.00	1.81%	175.50	2.05%
INTERGENIC	811.67	15.09%	899.00	13.96%	1,011.25	14.29%	1,199.25	14.00%
INTRON	221.67	4.12%	307.33	4.77%	351.75	4.97%	413.50	4.83%
SPLICE_SITE ACCEPTOR	4.67	0.09%	2.67	0.04%	7.25	0.10%	13.00	0.15%
SPLICE_SITE DONOR	2.50	0.05%	2.00	0.03%	2.67	0.04%	26.00	0.30%
SPLICE_SITE REGION	27.00	0.50%	41.33	0.64%	34.75	0.49%	68.50	0.80%
UPSTREAM	2,121.67	39.44%	2,546.33	39.54%	2,690.50	38.03%	3,381.75	39.47%
UTR_3_PRIME	72.33	1.34%	124.33	1.93%	132.50	1.87%	153.25	1.79%
UTR_5_PRIME	65.00	1.21%	91.67	1.42%	96.75	1.37%	102.25	1.19%

Table 17: Functional Predictions of Functional Changes Based on Genomic Variants Based on the functional prediction of SNPeff it used the genomic annotation file along with the type of change and location in the gene model to determine the location and type of change the would occur as a result of that variant.

INDELs

Allele frequency

Much in the same fashion as the SNPs, the INDELs need to be examined for what kind of variants are present. As mentioned above the majority of the variants are heterozygous. Based on the high percentages, it is no surprise that the allele

frequencies of the indels are more than 93% heterozygous. In examining the results of this part of the analysis, there does not seem to be any clear trend other than by overall numbers. Neither P14 nor *als3-3* show a consistent rate of change. Though the overall trend is consistent with the original trend of all variants for these samples.

There is also an interesting correlation between the totals of the indels of the P14 on 1.5 mM soak solution and those of the *als3-3* on 0.75 mM. This ties into growth phenotyping results observed in the lab using the *als3-3* genotype. *als3-3* hypersensitive phenotype to Al^{3+}, shows that on 0.75 mM gel soak media, the number of detectable Al associated variants of the P14 plants on 1.5 mM gel soak media. This present again that these genotypes are not comparable.

P14 Al Induced INDEL Allele Frequency				
Allele Freq	0.75 mM counts	0.75 mM %	1.5 mM counts	1.5 mM %
Homozygous	104	6.04	103	4.40
Heterozygous	1618	93.96	2236	95.60
Total	1722	100	2339	100

Table 18: P14 Al Induced INDEL Allele Frequency
The total amounts based on the non binary results using just those variants that fell in to the Al induced change category. Containing both the 0.75 mM and 1.5 mM treatments for comparison along with the corresponding percentage.

als3-3 Al Induced INDEL Allele Frequency				
Allele Freq	0.75 mM counts	0.75 mM %	1.5 mM counts	1.5 mM %
Homozygous	82	3.42	126	4.45
Heterozygous	2313	96.58	2703	95.55
Total	2395	100	2829	100

Table 19: *als3-3* Al Induced INDEL Allele Frequency
The total amounts based on the non binary results using just those variants that fell in to the Al induced change category. Containing both the 0.75 mM and 1.5 mM treatments for comparison along with the corresponding percentage.

Genomic Hotspots

The fact that the analysis can detect indels motivates the question of *is there any association between exposure to Al^{3+} and the location in the genome these INDELs are occuring*. If there are certain locations that have a high affinity for indels it could indicate possible fragile sites that correlate to Al^{3+} exposure. To test the possibility of genomic hotspots the INDELs were binned using 100 kb windows to asses the genomic impact based on location. The heat map of the chromosomes of Arabidopsis was generated, with blue as the lowest, yellow as 50% marker and red as the max values. The genomic binning of the hotspots includes all Al associated INDELs for each treatment and simply looks at overall distribution among the various genomic locations. In this case there are some regions of the genome that are either consistent between the two treatments.

Different from the SNPs however there are some regions that could be potentially showing a dose dependent response, where in low treatment the signal is only showing yellow but with the high treatment in the same are the signal shown turns to red. As mentioned previously however the genomic heat map with the binning results is only a preliminary visualization and requires further study to validate the results (Figures 28, 29).

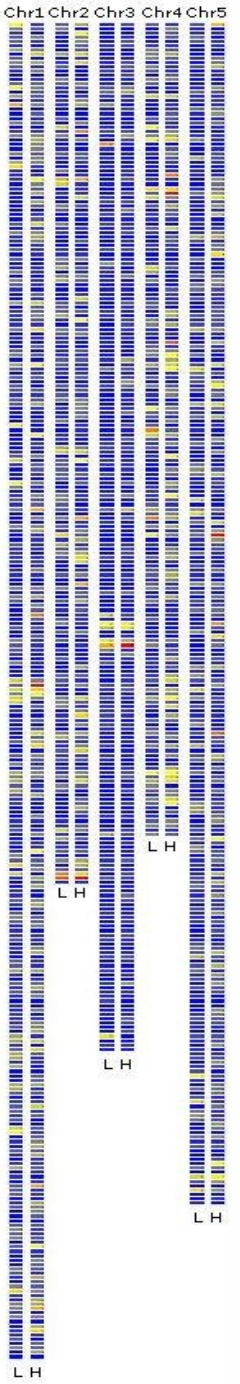


Figure 29: P14 Genomic Plot of INDELs in 100 kb Bins
Starting from the left, the heat map shows chromosomes 1-5 and the average rate of detected Aluminum INDELs across each chromosome. Those locations ranging from minimum of zero in the dark blue to the max values for both treatments in red for each chromosome. Each chromosome has two columns the left being the results of the 0.75 mM experiment (low treatment), and the right column the results of 1.5 mM experiment (high treatment)

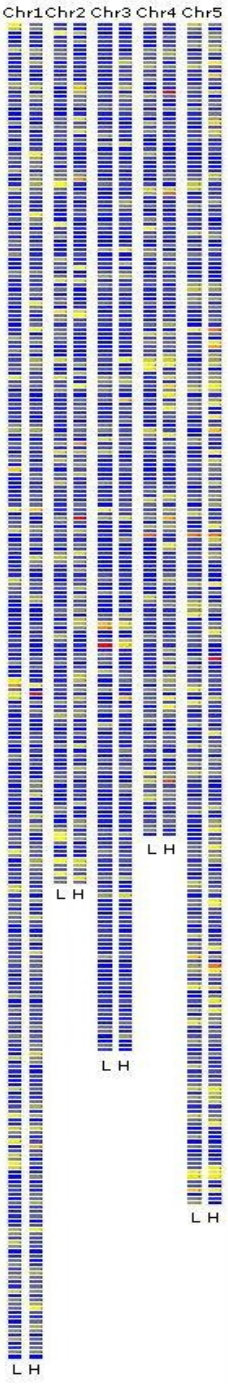


Figure 30: *als3-3* Genomic Plot of INDELs in 100 kbp Bins Starting from the left, the heat map shows chromosomes 1-5 and the total counts of all replicates of detected Aluminum INDELs across each chromosome. Those locations ranging from minimum of zero in the dark blue to the max values for both treatments in red for each chromosome. Each chromosome has two columns the left being the results of the 0.75 mM experiment (low treatment), and the right column the results of 1.5 mM experiment (high treatment)

Size distribution

Presented here is the overall size distribution of the wild type INDELs, followed by the size distribution, breaking them down by first insertions and then by deletions. This illustrates the range of what is detected by the variant caller and displayed as box plot that contains the combined data of each replicate done for each genotype. Box plots which are displayed by the median as the bar and the interquartile value shown by the box demonstrate the range of values contained within. Outliers are shown as dots on the graph that are separate from the box plots.

In the pursuit to see if a dose dependent pattern could be identified the Al dependent variant changes were compared for each genotype. It was of interest to see if any large scale patterns appeared when focusing on the two main types of INDELs independently. Upon initial inspection of the P14 samples breaking down the results in to insertions and deletions, the insertions appears more numerous. However in taking a closer look the 1 bp insertions skew the sizing of the graph, however the 2 bp insertions and deletions are comparable in order of magnitude. Using the overlap of the interquartile ranges of the box plots as rough estimate of what might be significant, it appears that after 3 bp INDELs in size, the interquartile ranges were almost equal. This would be tested further with ANOVA to verify, that only the INDELs 4 bp or smaller should be the focus of the rest of the study, with ability to expand this search if the results proved otherwise.

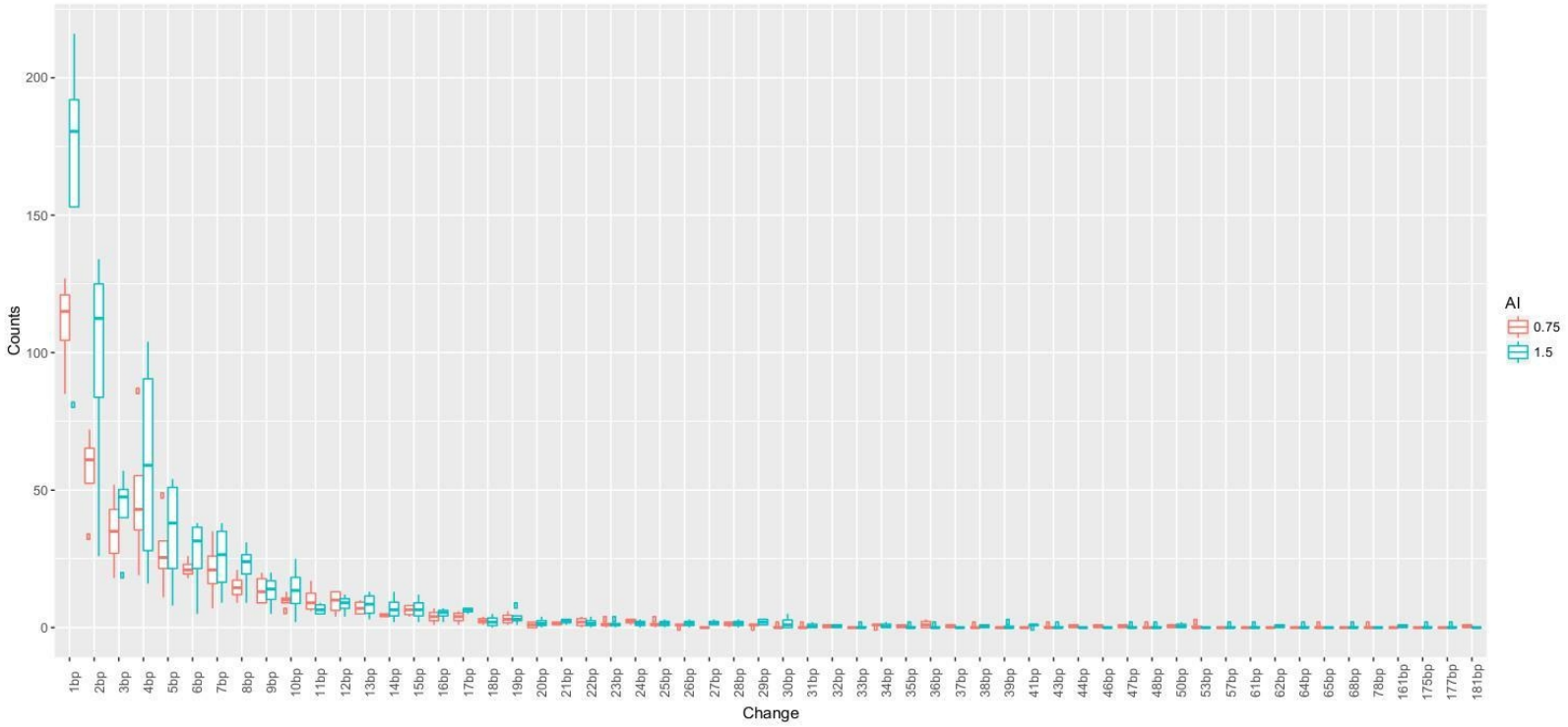


Figure 31: P14 Total INDEL distribution
Shown in the figure box plots representing the combined data from each rep of the P14 sequencing samples for those variants which were determined to be associated Al treatment. The 0.75 mM (low treatment) shown in blue and the 1.5 mM treatment shown in red. These INDELs and broken down by size (denoted by the horizontal axis) with the rates represented on the vertical axis. Outlier values are shown as dots seperate from the rest of the box and whiskers.

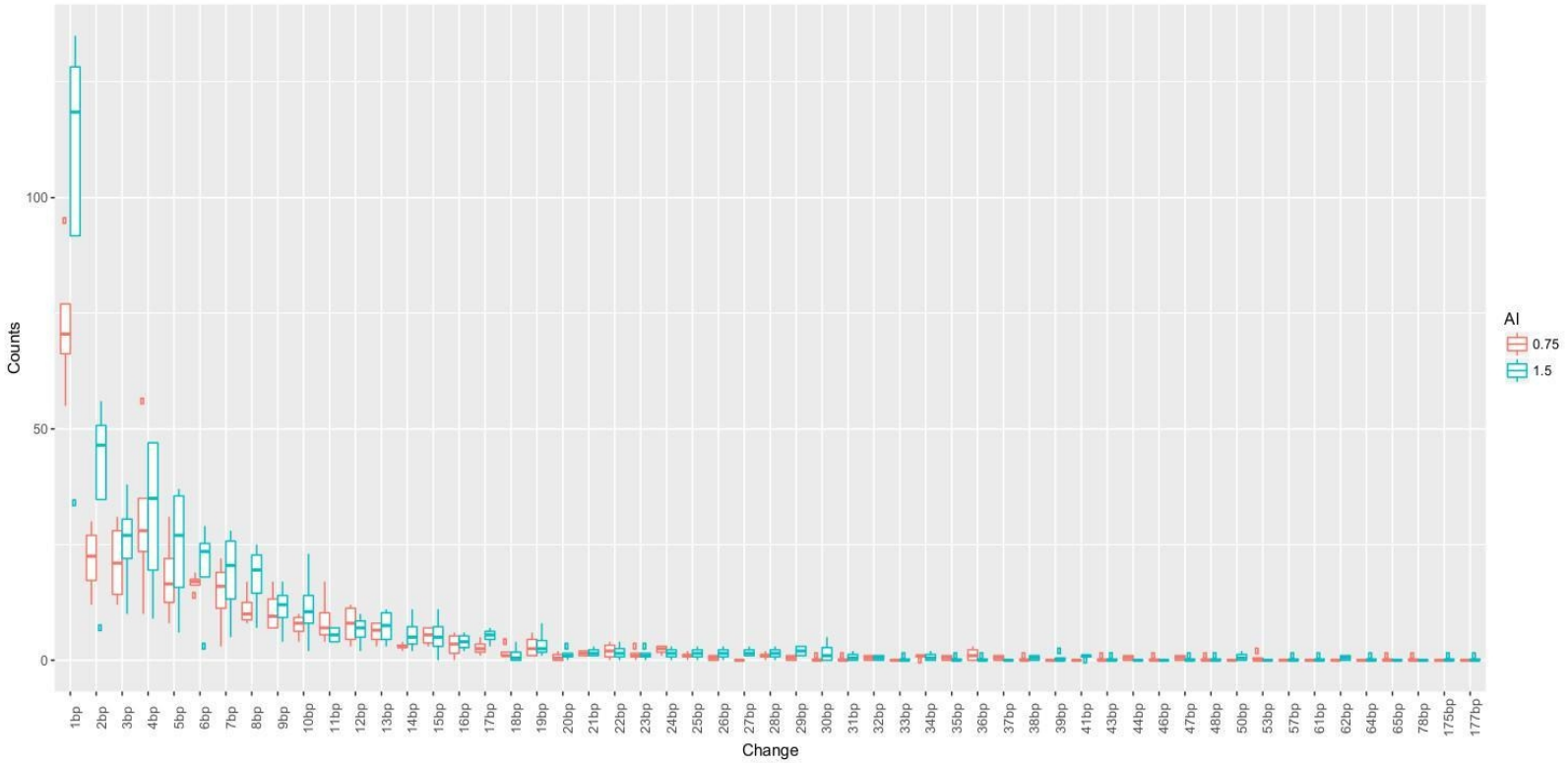


Figure 32: P14 Total Insertion Distribution
Shown in the figure box plots representing the combined data from each rep of the P14 sequencing samples for those variants which were determined to be associated Al treatment. The variants are a subset of the total INDELs and that just those that were called as leading to an insertion event. The 0.75 mM (low treatment) shown in blue and the 1.5 mM treatment shown in red. These INDELs and broken down by size (denoted by the horizontal axis) with the rates represented on the vertical axis. Outlier values are shown as dots seperate from the rest of the box and whiskers.

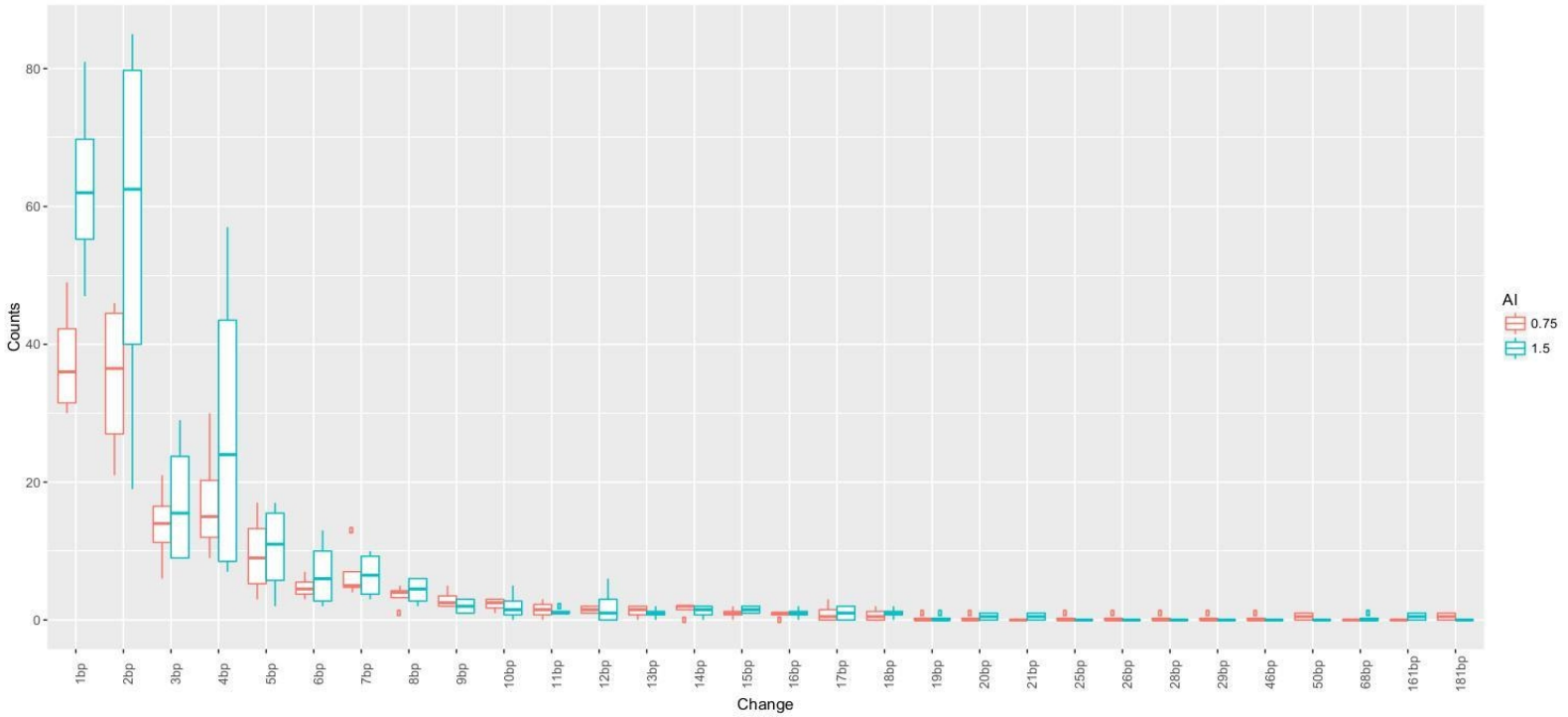


Figure 33: P14 Total Deletion Distribution
Shown in the figure box plots representing the combined data from each rep of the P14 sequencing samples for those variants which were determined to be associated Al treatment. The variants are a subset of the total INDELs and that just those that were called as leading to an deletion variant. The 0.75 mM (low treatment) shown in blue and the 1.5 mM treatment shown in red. These INDELs and broken down by size (denoted by the horizontal axis) with the counts represented on the vertical axis. Outlier values are shown as dots separate from the rest of the box and whiskers.

The *als3-3* total INDELs (Figure 33) look very similar to the P14 INDELs (Figure 30) in terms of distribution. What is noticeably different however is there is very little difference between the two treatments in *als3-3*. The median values show an increase but overall they do not show any sort major differences in the interquartile ranges. Here the data is again split into insertion and deletions to see if that would change the outlook of the overall trend.

Not surprisingly the trends remained the same. The data seems to suggest that *als3-3* due to its hypersensitivity has already reached some threshold of genomic damage, at least in terms of dose response. One additional pattern that can be seen in the overall INDEL distributions for both *als3-3* and P14 is a logarithmic trend decreasing from 1 bp indels towards a rate of zero with larger sized indels. Interestingly the 3 bp INDELs seem to defy the overall logarithmic trend in data. While this event happens in both genotypes it is much more noticeable in *als3-3* due to its increased counts of detected genomic variants associated with Al treatment. To be unbiased in the research the threshold for the P14 was used in determining the 4 bp cutoff to perform the ANOVA on.

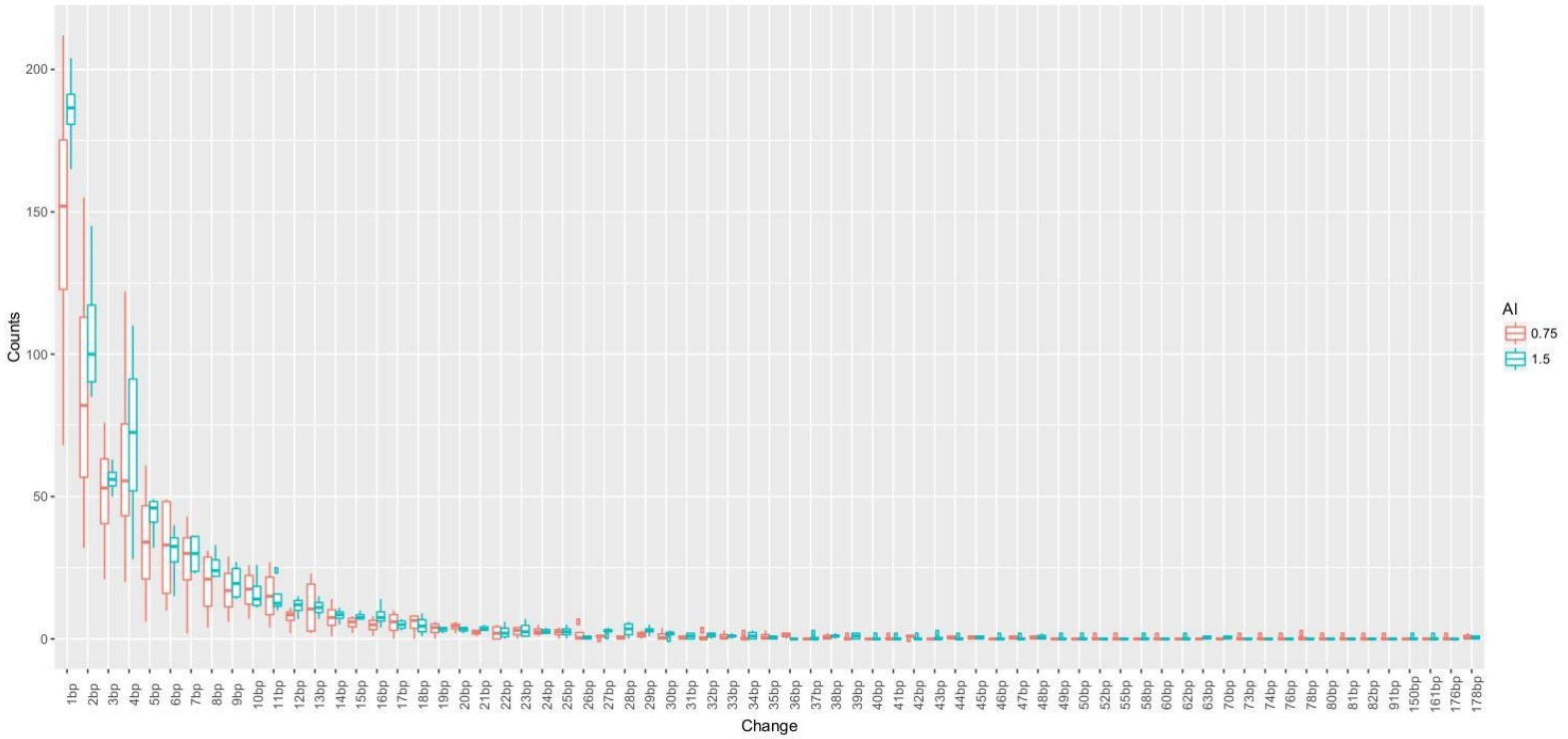


Figure 34: Total *als3-3* INDEL distribution
Shown in the figure box plots representing the combined data from each rep of the *als3-3* sequencing samples for those variants which were determined to be associated with Al treatment. The 0.75 mM (low treatment) shown in blue and the 1.5 mM treatment shown in red. These INDELs are broken down by size (denoted by the horizontal axis) with the counts represented on the vertical axis. Outlier values are shown as dots separate from the rest of the box and whiskers.

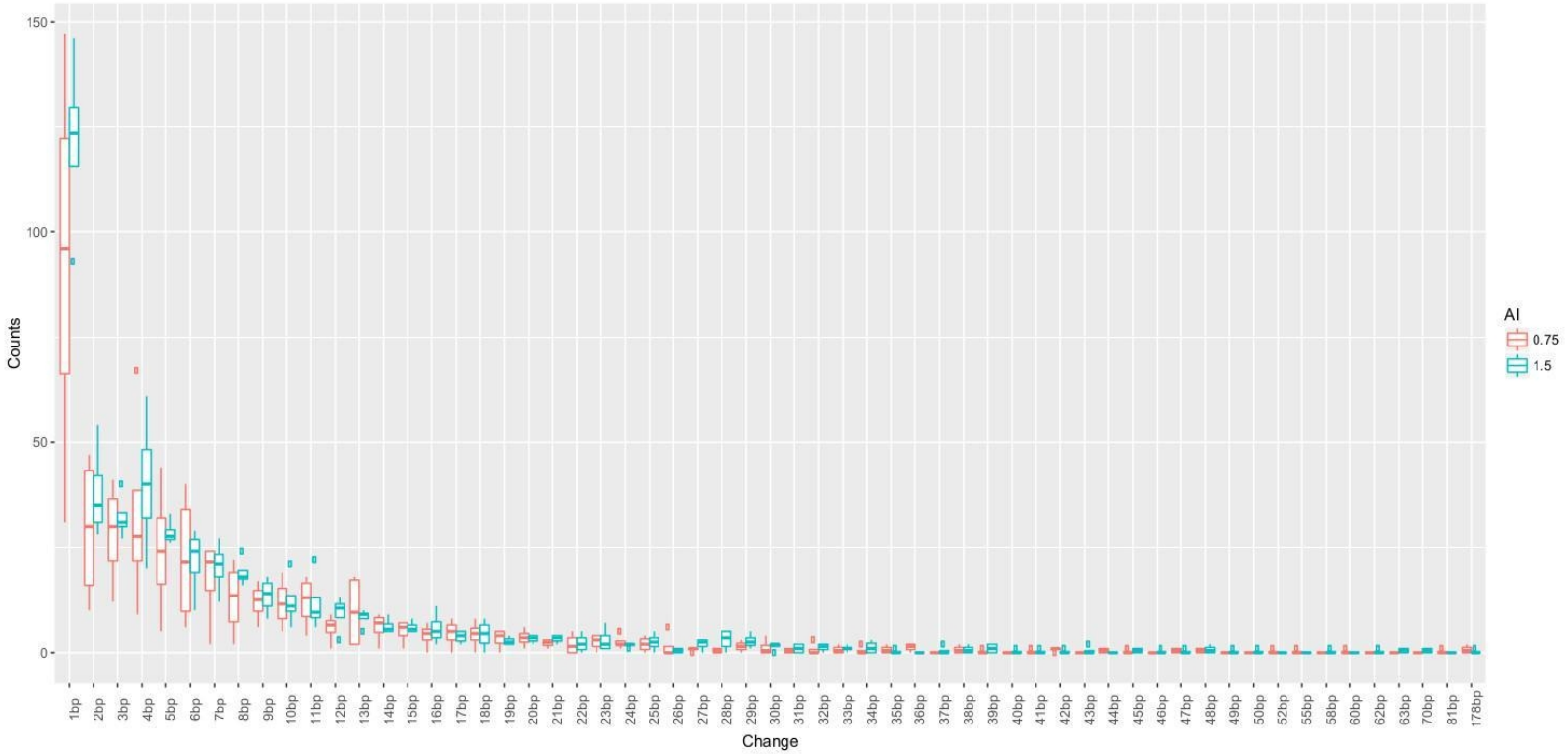


Figure 35: Total *als3-3* Insertion distribution
Shown in the figure box plots representing the combined data from each rep of the *als3-3* sequencing samples for those variants which were determined to be associated Al treatment. The variants are a subset of the variants that represent insertions and just those that were called as leading to an deletion variant. The 0.75 mM (low treatment) shown in blue and the 1.5 mM treatment shown in red. These INDELs and broken down by size (denoted by the horizontal axis) with the rates represented on the vertical axis. Outlier values are shown as dots seperate from the rest of the box and whiskers.

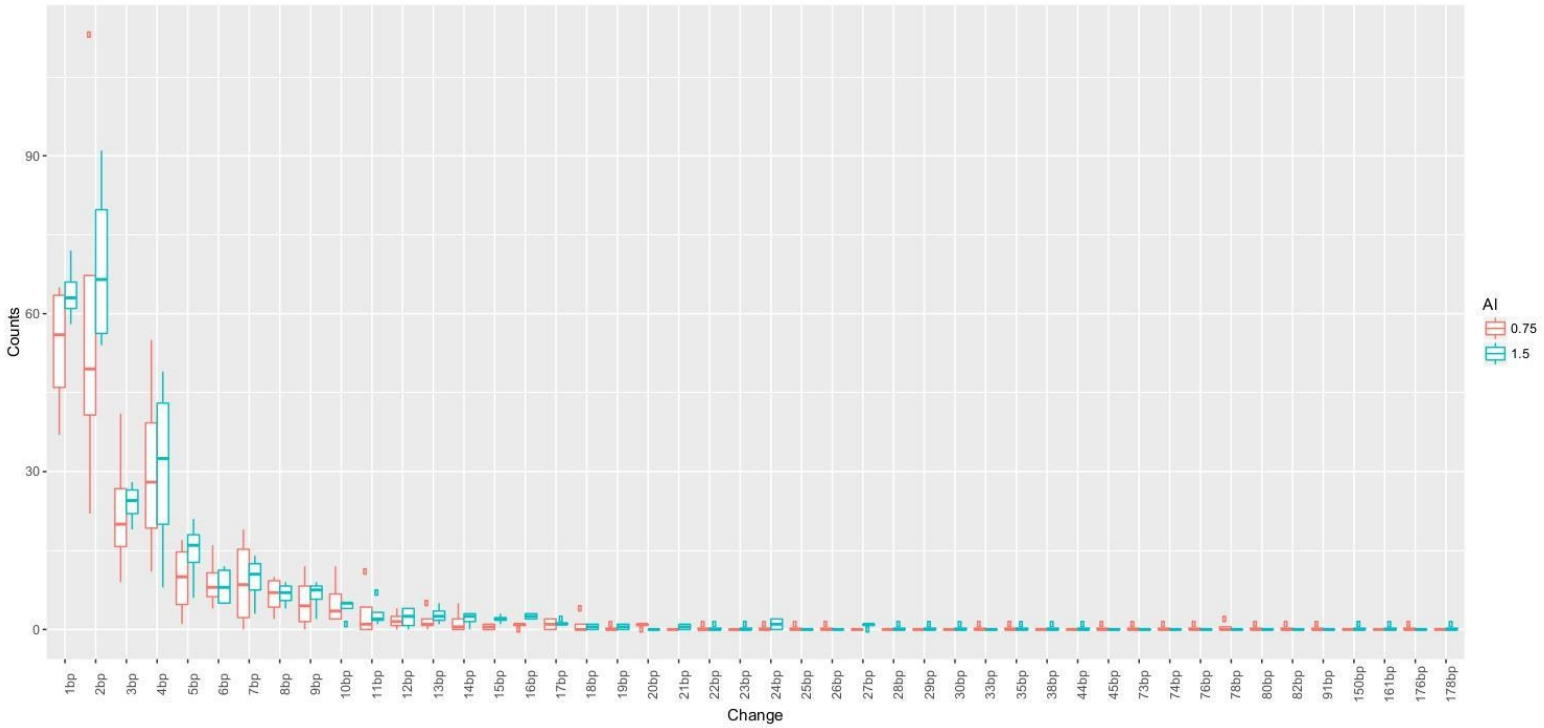


Figure 36: Total *als3-3* Deletion distribution
Shown in the figure box plots representing the combined data from each rep of the *als3-3* sequencing samples for those variants which were determined to be associated Al treatment. The variants are a subset of the INDELs determined to be deletions, and that just those that were called as leading to an deletion variant The 0.75 mM (low treatment) shown in blue and the 1.5 mM treatment shown in red. These INDELs are broken down by size (denoted by the horizontal axis) with the rates represented on the vertical axis. Outlier values are shown as dots seperate from the rest of the box and whiskers.

Small INDELs

The results from the analysis point to smaller indels being higher in frequency, that is where the study finds its focus to using just those INDELs that are 1-4 bp in size to look for dependent patterns in frequency and well as in the pattern of changes. 4 bp was chosen as the cut off for the total INDEL distribution as there the differences of the interquartile values of the counts of genomic variants occurring between dosage of the treatment in any INDELs larger than 4 bp were not large enough to see via the box plot (Figure 30). As previously mentioned in this section when looking at the inter-quartile values, 4bp is a good preliminary threshold value, which could be tested with the condition that if 4bp indels were found to be significant the threshold would need to be pushed further back to help avoid false negatives due to naive cutoffs.

As shown in figure 37 and 40, as the INDELs get larger, the difference in frequency with relation to the dose decreases. Based on these large difference shown deeper analysis was warranted to see if there was any change in frequency based on the actual pattern of the nucleotides being altered possibly by Al exposure. In an attempt to perform this analysis with minimal bias, ANOVA was used for each set of indels based on size. Then for those INDELS that were determined to be significant with respect the treatment factor were tested further for significance using the Poisson distribution in an attempt to identify which types of INDELs were significant.

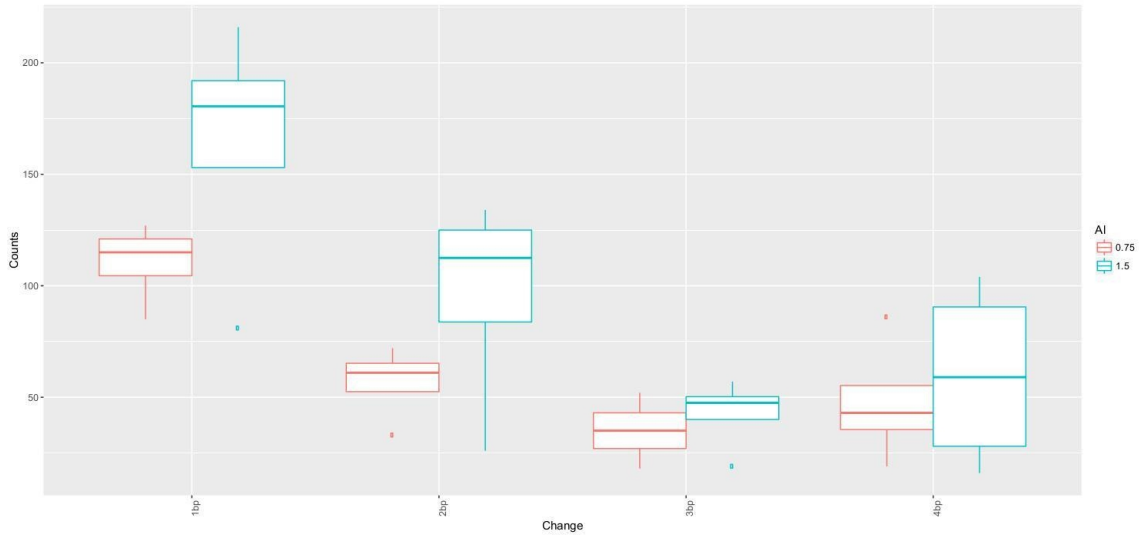


Figure 37: P14 INDELs 1-4 bp in Size
The INDELs shown are the same as those in the Figure X showing the total INDELs for P14, merely focused and resized on those indels 1-4 bp in size. The low treatment is red and the high treatment is shown in blue. With the size of the change as horizontal axis and the counts as the vertical axis.

ANOVA was performed for each Size of INDEL, for those INDELs of 1-4 bp identified as Al^{3+} associated those of the sizes 1bp and 2 bp were shown to be significant at $p < 0.001$, and the 2 bp INDELs showed a significant interaction between the Al^{3+} treatment and the types of changes at $p<0.01$. The 3 bp and 4 bp INDELs did not demonstrate significant changes in frequency associated with the Al treatment even though the differences between each type of indel (denoted by the “change row) themselves showed significant changes. However, as those changes were not dose dependent they were not considered for further analysis.

1 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	23	5337	232.06	16.84	< 2e-16*
Al treatment	1	243	243	17.63	4.68e-05*
type:Al	23	364	15.85	1.15	0.301
Residuals	144	1985	13.78		
2 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	71	1042.7	14.69	6.972	< 2e-16*
Al treatment	1	43.3	43.34	20.575	7.44e-06*
type:Al	71	227.7	3.21	1.522	0.00662*
Residuals	432	910	2.11		
3 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	94	503.9	5.36	9.351	<2e-16*
Al treatment	1	1.3	1.264	2.206	0.138
type:Al	94	47.9	0.509	0.888	0.759
Residuals	570	326.7	0.573		
4 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	146	1064.1	7.289	10.331	<2e-16*
Al treatment	1	1.9	1.878	2.663	0.103
type:Al	146	74.2	0.509	0.721	0.993
Residuals	882	622.3	0.705		

Table 20: ANOVA Results for P14 INDELs of 1-4 bp in Size
To provide statistical confirmation to the results seen in the box and whiskers plot for the INDELS 1-4 bp in size, each category of changes: 1 bp , 2bp ext. was tested by ANOVA to determine if there was significance to the change by means of the dose of Al^{3+} denoted in the table by the Al row. Using the concentration as a factor we could focus on just those categories that show a dose dependent change. While the 1-2 bp INDELs show significant changes that show relation to the factor of the Al^{3+} dose the 3-4 bp INDELs were excluded from further analysis because while the changes show significance they are not related to the treatment.

For those variants identified as 1 bp and 2 bp INDELs using the Poisson distribution with a p-value cutoff of $p < 0.01$ (denoted as one or more * next to the p-values) to determine if the identified variant is significant. For notation purposes, the bases used from the TAIR10 reference genome are denoted first followed by the asterisk and the variant called. For example, A*AC signifies an insertion where the reference genome says that position has an “A” but the sequencing results show that an “AC” is present in that position, signifying that there is a 1 bp insertion of “C” after the “A”. Conversely for a deletion AC*A would mean that a C was deleted according to the sequencing results.

In looking at overall types of these variants it is apparent that many of these types of INDELs are reciprocals of one another. Meaning that a one type of deletion is matched by an insertion that seems to be the opposite of the deletion, this has many possibilities however that this analysis does not explore. However this is not the case for all of these changes. To further assess the consequences for these changes and those of the *als3-3* genotype SNPeff [98] is used once more to assess the possible consequences of these changes in a more broad scope.

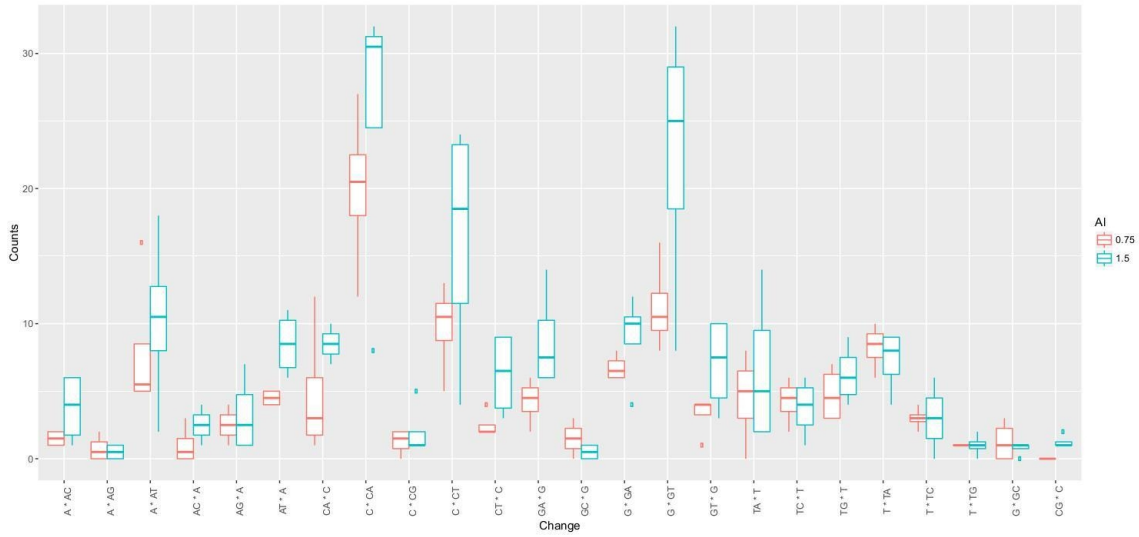


Figure 38: P14 INDELs 1 bp in Size

Box and whiskers plot visualizing the results of the 1 bp INDELs based on the types of changes found from the variant calling and filtering. The visualization provided a double check of the statistical analysis as to which types of INDELs are significant in a dose dependent manner.

INDEL Type	Estimate	Std. Error	Z Value	p Value
Change A*AC	0.74647	0.2213	3.373	0.000743*
Change A*AG	-1.43508	0.49761	-2.884	0.003927*
Change A*AT	1.24594	0.24762	5.032	0.000000486*
Change AC*A	-0.40547	0.34503	-1.175	0.239935
Change AG*A	0.09097	0.30182	0.301	0.763104
Change AT*A	0.90672	0.25855	3.507	0.000453*
Change CA*C	0.92577	0.25785	3.59	0.00033*
Change C*CA	2.15397	0.23053	9.344	< 2E-16*
Change C*CG	-0.47957	0.35291	-1.359	0.174169
Change C*CT	1.59987	0.23924	6.687	2.27E-11*
Change CT*C	0.51083	0.27603	1.851	0.064221
Change GA*G	0.90672	0.25855	3.507	0.000453*
Change GC*G	-0.96508	0.41547	-2.323	0.020186*
Change G*GA	1.09861	0.25198	4.36	0.000013*
Change G*GT	1.86075	0.23458	7.932	2.15E-15*
Change GT*G	0.66905	0.26835	2.493	0.012658*
Change TA*T	0.73967	0.26523	2.789	0.005291*
Change TC*T	0.42121	0.28084	1.5	0.133652
Change TG*T	0.73967	0.26523	2.789	0.005291*
Change T*TA	1.08261	0.25248	4.288	0.000018*
Change T*TC	0.13353	0.29881	0.447	0.654961
Change T*TG	-0.96508	0.41547	-2.323	0.020188*
Change G*GC	-0.96508	0.41547	-2.323	0.020187*
Change CG*C	-1.43508	0.49761	-2.884	0.003927*
Al treatment 1.5 mM	0.3979	0.0615	6.47	9.81E-11*

Table 21: Results of Testing 1 bp P14 Variants Using Poisson Distribution
Results of testing for which type of indels were significant using the Poisson distribution those types of indels that are marked with one or more stars are significant at p <0.5, there does not appear to be any strict patterns around the types of significant INDELs. Many of the Deletions are matched by the reciprocal insertion. In referring back to the corresponding box and whiskers plot, additional inference can be made as some of these changes while significant with relation to dose are inverse in rate relative to dose.

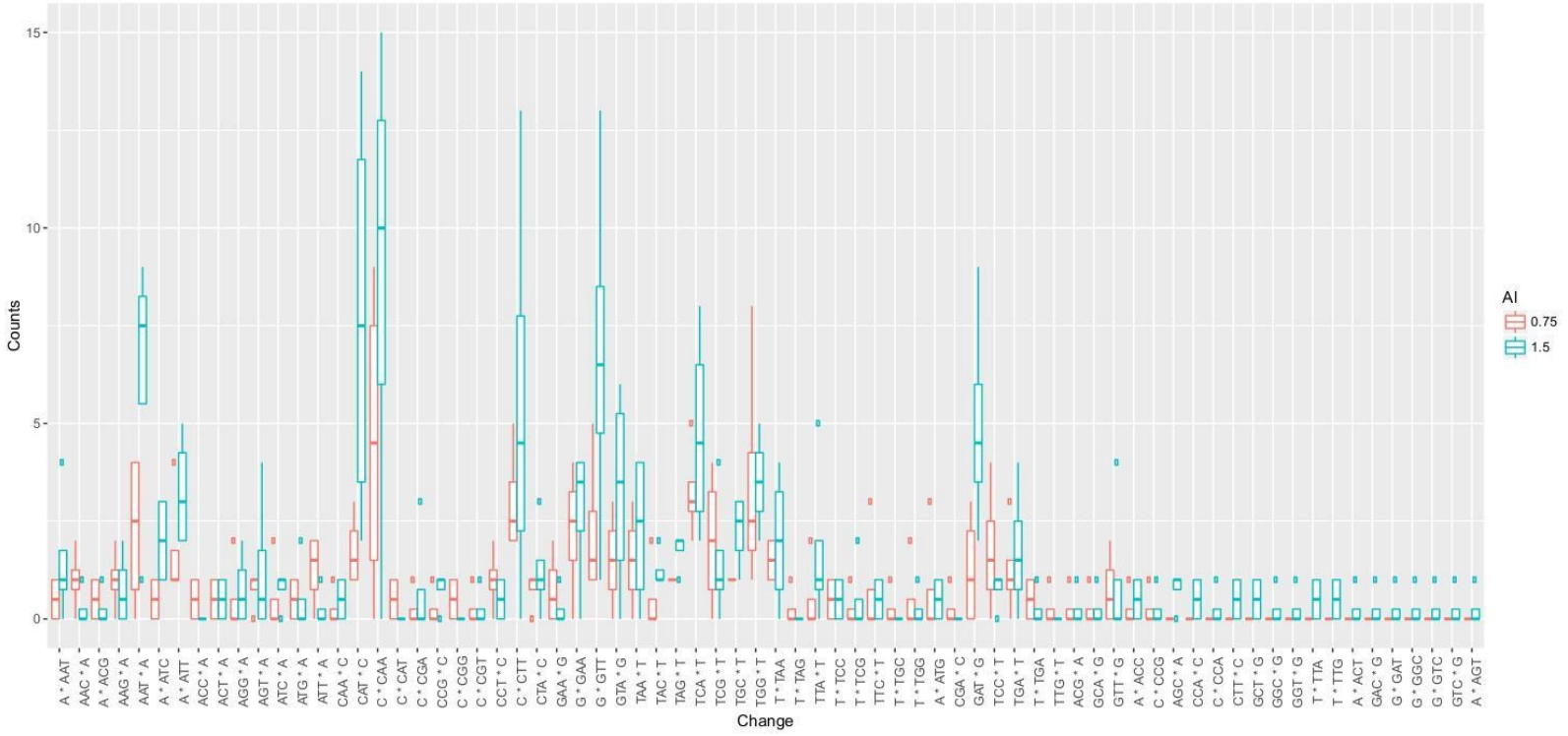


Figure 39: P14 INDELs 2 bp in Size
Box and whiskers plot visualizing the results of the 2 bp INDELs based on the types of changes found from the variant calling and filtering. The visualization provided a double check of the statistical analysis as to which types of INDELs are significant in a dose dependent manner.

INDEL Type	Estimate	Std Error	Z Value	p Value
Change A*AAT	-2.99E-01	3.58E-01	-0.835	0.40346
Change AAC*A	-4.70E-01	5.70E-01	-0.824	0.409689
Change A*ACG	-9.81E-01	6.77E-01	-1.449	0.147399
Change AAG*A	-1.34E-01	5.18E-01	-0.258	0.796401
Change AAT*A	1.45E+00	3.93E-01	3.682	0.000231*
Change A*ATC	2.23E-01	4.74E-01	0.47	0.638049
Change A*ATT	9.16E-01	4.18E-01	2.19	0.028499*
Change ACC*A	-1.39E+00	7.91E-01	-1.754	0.079509
Change ACT*A	-6.93E-01	6.12E-01	-1.132	0.257675
Change AGG*A	-4.70E-01	5.70E-01	-0.824	0.409689
Change AGT*A	-4.62E-16	5.00E-01	0	1
Change ATC*A	-4.70E-01	5.70E-01	-0.824	0.409689
Change ATG*A	-6.93E-01	6.12E-01	-1.132	0.257675
Change ATT*A	-2.88E-01	5.40E-01	-0.533	0.594253
Change CAA*C	-9.81E-01	6.77E-01	-1.449	0.147399
Change CAT*C	1.56E+00	3.89E-01	4.006	0.0000619*
Change C*CAA	1.89E+00	3.79E-01	4.985	0.000000619*
Change C*CAT	-1.39E+00	7.91E-01	-1.754	0.079509
Change C*CGA	-6.93E-01	6.12E-01	-1.132	0.257674
Change CCG*C	-6.93E-01	6.12E-01	-1.132	0.257675
Change C*CGG	-1.39E+00	7.91E-01	-1.754	0.079509
Change C*CGT	-1.39E+00	7.91E-01	-1.754	0.07951
Change CCT*C	-2.88E-01	5.40E-01	-0.533	0.594253
Change C*CTT	1.45E+00	3.93E-01	3.682	0.000231*
Change CTA*C	-3.70E-15	5.00E-01	0	1
Change GAA*G	-6.93E-01	6.12E-01	-1.132	0.257675
Change G*GAA	9.16E-01	4.18E-01	2.19	0.028499*
Change G*GTT	1.50E+00	3.91E-01	3.848	0.000119*
Change GTA*G	8.65E-01	4.22E-01	2.052	0.040134*
Change TAA*T	6.29E-01	4.38E-01	1.436	0.151047
Change TAC*T	-1.34E-01	5.18E-01	-0.258	0.796401

Change TAG*T	3.19E-01	4.65E-01	0.685	0.493125
Change TCA*T	1.39E+00	3.95E-01	3.507	0.000453*
Change TCG*T	5.60E-01	4.43E-01	1.263	0.20671
Change TGC*T	4.86E-01	4.49E-01	1.08	0.279943
Change TGG*T	1.25E+00	4.01E-01	3.125	0.001778*
Change T*TAA	5.60E-01	4.43E-01	1.263	0.20671
Change T*TAG	-2.08E+00	1.06E+00	-1.961	0.049935*
Change TTA*T	1.18E-01	4.86E-01	0.242	0.808474
Change T*TCC	-6.93E-01	6.12E-01	-1.132	0.257675
Change T*TCG	-9.81E-01	6.77E-01	-1.449	0.147399
Change TTC*T	-4.70E-01	5.70E-01	-0.824	0.409689
Change T*TGC	-2.08E+00	1.06E+00	-1.961	0.049935*
Change T*TGG	-9.81E-01	6.77E-01	-1.449	0.147399
Change A*ATG	-4.70E-01	5.70E-01	-0.824	0.409689
Change CGA*C	-2.08E+00	1.06E+00	-1.961	0.049935*
Change GAT*G	1.14E+00	4.06E-01	2.805	0.00503*
Change TCC*T	2.23E-01	4.74E-01	0.47	0.638049
Change TGA*T	4.06E-01	4.56E-01	0.888	0.374364
Change T*TGA	-9.81E-01	6.77E-01	-1.449	0.147399
Change TTG*T	-2.08E+00	1.06E+00	-1.961	0.049935*
Change ACG*A	-1.39E+00	7.91E-01	-1.754	0.07951
Change GCA*G	-1.39E+00	7.91E-01	-1.754	0.07951
Change GTT*G	-1.34E-01	5.18E-01	-0.258	0.796401
Change A*ACC	-9.81E-01	6.77E-01	-1.449	0.147399
Change C*CCG	-1.39E+00	7.91E-01	-1.754	0.07951
Change AGC*A	-9.81E-01	6.77E-01	-1.449	0.147399
Change CCA*C	-1.39E+00	7.91E-01	-1.754	0.07951
Change C*CCA	-2.08E+00	1.06E+00	-1.961	0.049935*
Change CTT*C	-1.39E+00	7.91E-01	-1.754	0.07951
Change GCT*G	-1.39E+00	7.91E-01	-1.754	0.07951
Change GGC*G	-2.08E+00	1.06E+00	-1.961	0.049935*
Change GGT*G	-2.08E+00	1.06E+00	-1.961	0.049935*

Change T*TTA	-1.39E+00	7.91E-01	-1.754	0.07951
Change T*TTG	-1.39E+00	7.91E-01	-1.754	0.07951
Change A*ACT	-2.08E+00	1.06E+00	-1.961	0.049935*
Change GAC*G	-2.08E+00	1.06E+00	-1.961	0.049935*
Change G*GAT	-2.08E+00	1.06E+00	-1.961	0.049935*
Change G*GGC	-2.08E+00	1.06E+00	-1.961	0.049935*
Change G*GTC	-2.08E+00	1.06E+00	-1.961	0.049935*
Change GTC*G	-2.08E+00	1.06E+00	-1.961	0.049935*
Change A*AGT	-2.08E+00	1.06E+00	-1.961	0.049935*
Al 1.5	5.28E-01	8.37E-02	6.313	2.73e-10*

Table 22: Results of Testing 2 bp P14 Variants Using Poisson Distribution
Results of testing for which type of indels were significant when comparing the low treatment to the high treatment using the Poisson distribution those types of indels that are marked with one or more stars are significant at p <0.5, there appears to be a majority of AT deletions and homopolymer insertions of T and A repeats. In referring back to the corresponding box and whiskers plot, additional inference can be made as some of these changes while significant with relation to dose are are inverse in rate relative to the dose.

The same analysis which was performed for P14 was done using the same process for the *als3-3* samples. This meant starting with the INDELs less than 5 bp in size and performing ANOVA to test for which sizes of INDELs should be further analyzed. As shown in the table below again the 1-2 bp INDELs showed significance while 3-4 bp INDELs did not.

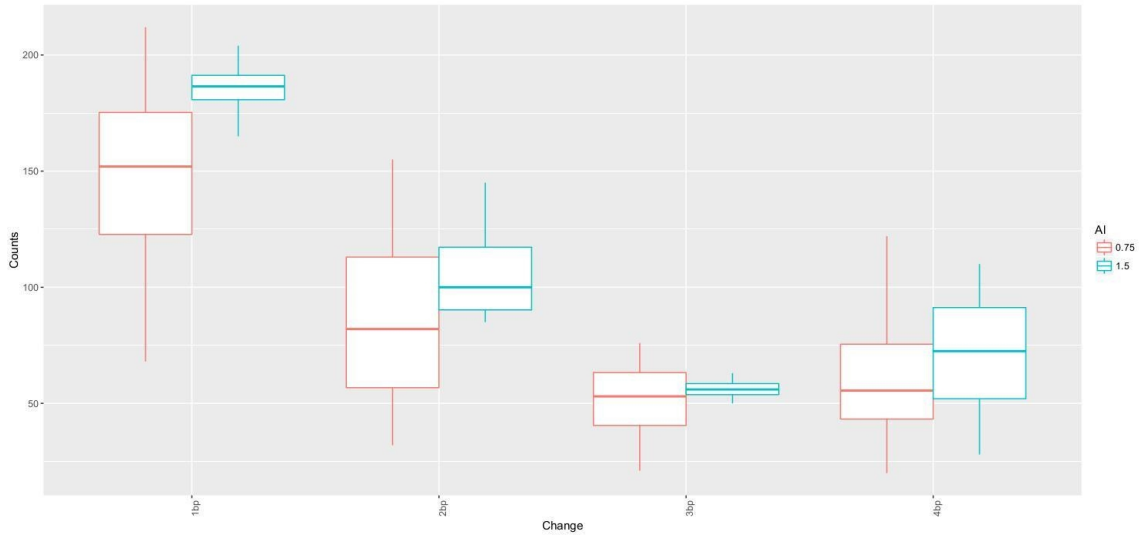


Figure 40: *asl3-3* INDELs 1-4 bp in Size
The INDELs shown are the same as those in the Figure X showing the total INDELs for *als3-3*, merely focused and resized on those indels 1-4 bp in size. The low treatment (0.75 mM) is red and the high treatment (1.5 mM) is shown in blue. With the size of the change as horizontal axis and the counts as the vertical axis. The bars represent the distribution of values with boxed indicating the interquartile values and the line in the box the median of the values.

1 bp	Degrees of freedom	Sum of squares	Mean Squared	F value	P value
INDEL type	23	9107	395.9	20.407	<2e-16*
Al treatment	1	130	130	6.701	0.0106*
type:Al	23	721	31.4	1.617	0.0473*
Residuals	144	2794	19.4		
2 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	75	1763.6	23.515	8.919	<2e-16*
Al treatment	1	10.3	10.265	3.893	0.0491*
type:Al	75	197.6	2.635	0.999	0.4846
Residuals	456	1202.3	2.637		
3 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	111	901.1	8.118	14.45	<2e-16*
Al treatment	1	0.5	0.54	0.962	0.327
type:Al	111	52.5	0.473	0.841	0.871
Residuals	672	377.5	0.562		
4 bp	**Degrees of freedom**	**Sum of squares**	**Mean Squared**	**F value**	**P value**
INDEL type	154	1613.3	10.476	9.792	<2e-16*
Al treatment	1	0.7	0.726	0.678	0.41
type:Al	154	51.3	0.333	0.311	1
Residuals	930	995	1.07		

Table 23: ANOVA Results for *als3-3* INDELs of 1-4 bp in Size
To provide statistical confirmation to the results seen in the box and whiskers plot for the INDELS 1-4 bp in size, each category of changes: 1 bp , 2bp ext. was tested by ANOVA to determine if there was significance to the change by means of the dose of Al^{3+} denoted in the table by the Al row. Using the concentration as a factor we could focus on just those categories that show a dose dependent change. While the 1-2 bp INDELs show significant changes that show relation to the factor of the Al^{3+} dose the 3-4 bp INDELs were excluded from further analysis because while the changes show significance they are not related to the treatment.

After using the ANOVA to further guide the analysis, 1-2 bp INDELs were analyzed to determine which variants were significant via the Poisson distribution. The ANOVA does not find the changes in 1-2 bp to be at the same level of significance as those from the P14 sample the best hypothesis for this would be overall increase in AI associated variants that were detected in this genotype.

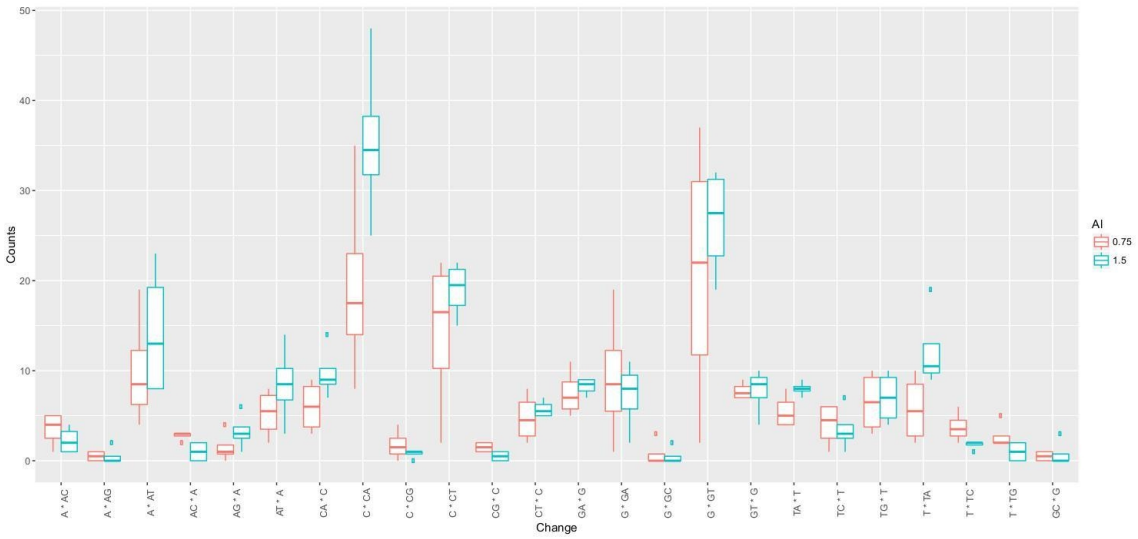


Figure 41: *als3-3* 1 bp Box and Whiskers Plot
Box and whiskers plot visualizing the results of the 1 bp INDELs based on the types of changes found from the variant calling and filtering. The visualization provided a double check of the statistical analysis as to which types of INDELs are significant in a dose dependent manner.

INDEL Type	Estimate	Std. Error	z value	p value
Change A * AC	0.92918	0.2108	4.408	0.0000104*
Change A * AG	-1.7492	0.54174	-3.229	0.001243*
Change A * AT	1.43922	0.23192	6.206	5.45E-10*
Change AC * A	-0.42744	0.33188	-1.288	0.197765
Change AG * A	-0.19106	0.31002	-0.616	0.537713
Change AT * A	0.87184	0.24831	3.511	0.000446*
Change CA * C	1.00764	0.24362	4.136	3.53E-05*
Change C * CA	2.25813	0.21914	10.304	< 2E-16*
Change C * CG	-0.83291	0.37879	-2.199	0.027885*
Change C * CT	1.75485	0.22583	7.771	7.80E-15*
Change CG * C	-1.05605	0.41046	-2.573	0.010087*
Change CT * C	0.60218	0.2594	2.321	0.020264*
Change GA * G	1.00764	0.24362	4.136	3.53E-05*
Change G * GA	1.05416	0.24214	4.354	1.34E-05*
Change G * GC	-1.52606	0.49343	-3.093	0.001983*
Change G * GT	2.10625	0.22084	9.538	< 2E-16*
Change GT * G	0.99164	0.24415	4.062	4.87E-05*
Change TA * T	0.85349	0.24899	3.428	0.000609*
Change TC * T	0.2657	0.27715	0.959	0.337709
Change TG * T	0.85349	0.24899	3.428	0.000609*
Change T * TA	1.14117	0.23951	4.765	1.89E-06*
Change T * TC	-0.04445	0.29822	-0.149	0.881507
Change T * TG	-0.42744	0.33188	-1.288	0.197765
Change GC * G	-1.52606	0.49344	-3.093	0.001983*
Al 1.5	0.23945	0.05532	4.329	1.50E-05*

Table 24: Results of Testing 1 bp *als3-3* Variants Using Poisson Distribution
Results of testing for which type of indels were significant using the Poisson distribution those types of indels that are marked with one or more stars are significant at p <0.5, the 1 bp INDELs do not appear to have any district trend other than certain INDELs had higher significance. In referring back to the corresponding box and whiskers plot, additional inference can be made as some of these changes while significant with relation to dose are are inverse in rate relative to the dose.

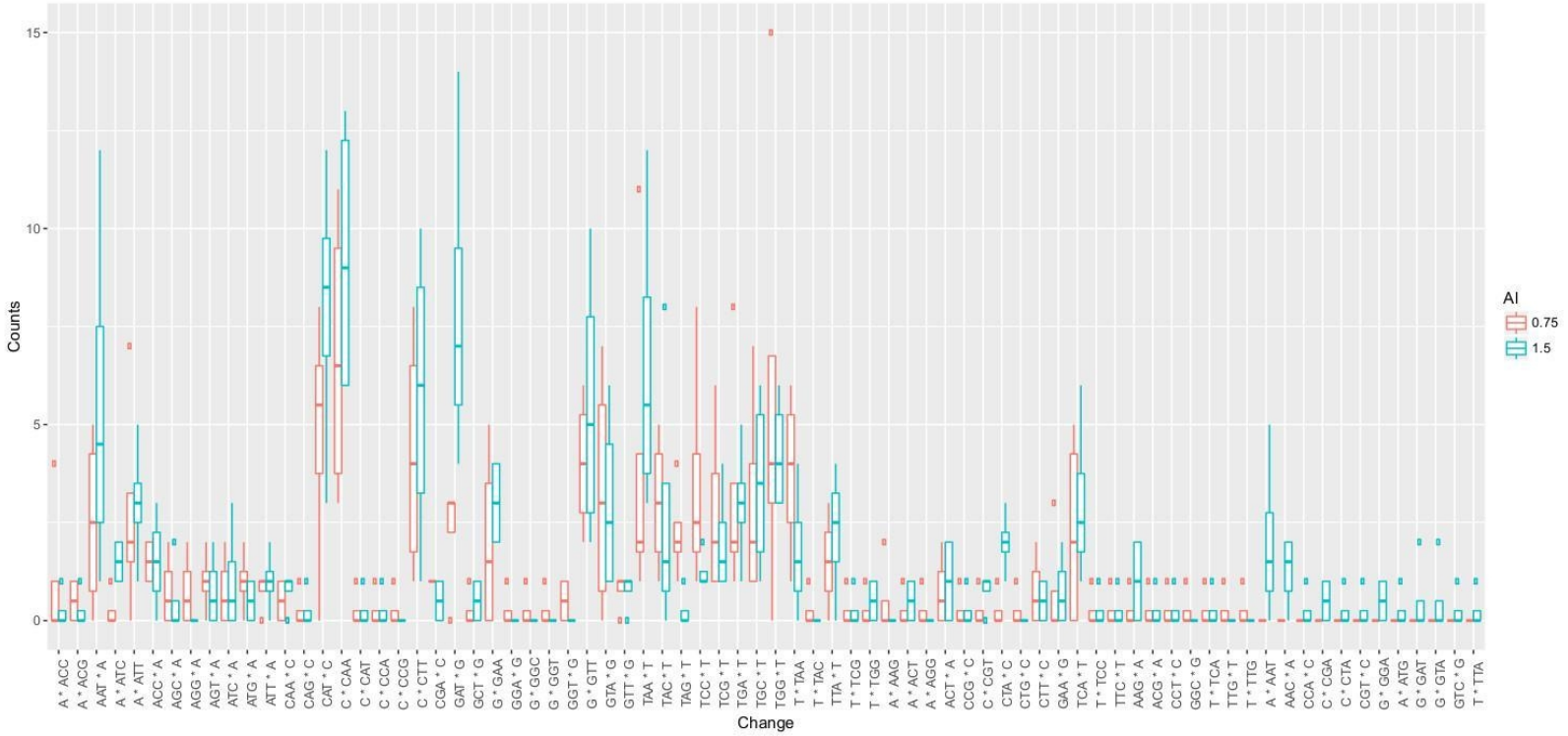


Figure 42: *als3-3* 2 bp Box and Whiskers Plot
Box and whiskers plot visualizing the results of the 2 bp INDELs based on the types of changes found from the variant calling and filtering. The visualization provided a double check of the statistical analysis as to which types of INDELs are significant in a dose dependent manner.

INDEL Type	Estimate	Std. Error	z Value	p Value
Change A*ACC	-5.77E-01	4.49E-01	-1.284	0.199003
Change A*ACG	-5.11E-01	7.30E-01	-0.699	0.484254
Change AAT*A	1.86E+00	4.81E-01	3.86	0.000113*
Change A*ATC	3.37E-01	5.86E-01	0.575	0.565537
Change A*ATT	1.53E+00	4.93E-01	3.093	0.001983*
Change ACC*A	8.76E-01	5.32E-01	1.645	0.100027
Change AGC*A	-2.70E-11	6.33E-01	0	1
Change AGG*A	-5.11E-01	7.30E-01	-0.699	0.484253
Change AGT*A	3.37E-01	5.86E-01	0.575	0.565537
Change ATC*A	3.37E-01	5.86E-01	0.575	0.565537
Change ATG*A	1.82E-01	6.06E-01	0.301	0.763342
Change ATT*A	3.37E-01	5.86E-01	0.575	0.565537
Change CAA*C	-2.70E-11	6.33E-01	0	1
Change CAG*C	-9.16E-01	8.37E-01	-1.095	0.273439
Change CAT*C	2.32E+00	4.69E-01	4.956	7.20E-07*
Change C*CAA	2.55E+00	4.64E-01	5.49	4.01E-08*
Change C*CAT	-9.16E-01	8.37E-01	-1.095	0.273439
Change C*CCA	-9.16E-01	8.37E-01	-1.095	0.273439
Change C*CCG	-1.61E+00	1.10E+00	-1.469	0.141776
Change C*CTT	2.08E+00	4.74E-01	4.384	1.17E-05*

Change CGA*C	1.82E-01	6.06E-01	0.301	0.763342
Change GAT*G	2.10E+00	4.74E-01	4.442	8.92E-06*
Change GCT*G	-5.11E-01	7.30E-01	-0.699	0.484254
Change G*GAA	1.39E+00	5.00E-01	2.773	0.005561*
Change GGA*G	-1.61E+00	1.10E+00	-1.469	0.141776
Change G*GGC	-1.61E+00	1.10E+00	-1.469	0.141776
Change G*GGT	-1.61E+00	1.10E+00	-1.469	0.141776
Change GGT*G	-9.16E-01	8.37E-01	-1.095	0.273439
Change G*GTT	2.03E+00	4.76E-01	4.263	2.01E-05*
Change GTA*G	1.61E+00	4.90E-01	3.285	0.001019*
Change GTT*G	1.82E-01	6.06E-01	0.301	0.763342
Change TAA*T	2.13E+00	4.73E-01	4.499	6.84E-06*
Change TAC*T	1.53E+00	4.93E-01	3.093	0.001983*
Change TAG*T	6.93E-01	5.48E-01	1.266	0.205688
Change TCC*T	1.34E+00	5.03E-01	2.656	0.007906*
Change TCG*T	1.34E+00	5.03E-01	2.656	0.007906*
Change TGA*T	1.61E+00	4.90E-01	3.285	0.001019*
Change TGC*T	1.65E+00	4.88E-01	3.376	0.000735*
Change TGG*T	2.08E+00	4.74E-01	4.384	1.17E-05*
Change T*TAA	1.48E+00	4.95E-01	2.991	0.002785*
Change T*TAC	-1.61E+00	1.10E+00	-1.469	0.141776

Change TTA*T	1.10E+00	5.16E-01	2.127	0.033382*
Change T*TCG	-9.16E-01	8.37E-01	-1.095	0.273439
Change T*TGG	-5.11E-01	7.30E-01	-0.699	0.484254
Change A*AAG	-9.16E-01	8.37E-01	-1.095	0.273438
Change A*ACT	-5.11E-01	7.30E-01	-0.699	0.484254
Change A*AGG	-1.61E+00	1.10E+00	-1.469	0.141776
Change ACT*A	3.37E-01	5.86E-01	0.575	0.565537
Change CCG*C	-9.16E-01	8.37E-01	-1.095	0.273439
Change C*CGT	-2.23E-01	6.71E-01	-0.333	0.739404
Change CTA*C	5.88E-01	5.58E-01	1.054	0.291969
Change CTG*C	-1.61E+00	1.10E+00	-1.469	0.141776
Change CTT*C	-2.70E-11	6.33E-01	0	1
Change GAA*G	1.82E-01	6.06E-01	0.301	0.763342
Change TCA*T	1.44E+00	4.98E-01	2.884	0.003927*
Change T*TCC	-9.16E-01	8.37E-01	-1.095	0.273439
Change TTC*T	-9.16E-01	8.37E-01	-1.095	0.273439
Change AAG*A	-2.70E-11	6.33E-01	0	1
Change ACG*A	-9.16E-01	8.37E-01	-1.095	0.273439
Change CCT*C	-9.16E-01	8.37E-01	-1.095	0.273439
Change GGC*G	-1.61E+00	1.10E+00	-1.469	0.141776
Change T*TCA	-9.16E-01	8.37E-01	-1.095	0.273439

Change TTG*T	-1.61E+00	1.10E+00	-1.469	0.141776
Change T*TTG	-1.61E+00	1.10E+00	-1.469	0.141776
Change A*AAT	4.70E-01	5.70E-01	0.824	0.409688
Change AAC*A	-2.70E-11	6.33E-01	0	1
Change CCA*C	-1.61E+00	1.10E+00	-1.469	0.141776
Change C*CGA	-9.16E-01	8.37E-01	-1.095	0.273439
Change C*CTA	-1.61E+00	1.10E+00	-1.469	0.141776
Change CGT*C	-1.61E+00	1.10E+00	-1.469	0.141776
Change G*GGA	-9.16E-01	8.37E-01	-1.095	0.273439
Change A*ATG	-1.61E+00	1.10E+00	-1.469	0.141776
Change G*GAT	-9.16E-01	8.37E-01	-1.095	0.273438
Change G*GTA	-9.16E-01	8.37E-01	-1.095	0.273438
Change GTC*G	-1.61E+00	1.10E+00	-1.469	0.141776
Change T*TTA	-1.61E+00	1.10E+00	-1.469	0.141776
Al 1.5 mM	2.03E-01	7.19E-02	2.822	0.004773*

Table 25: Results of Testing 2 bp *als3-3* Variants Using Poisson Distribution
Results of testing for which type of indels were significant using the Poisson distribution across the various types of indels that are marked with one or more stars are significant at p <0.5, there to be a majority of AT deletions and homopolymer insertions of T and A repeats. In referring back to the corresponding box and whiskers plot, additional inference can be made as some of these changes while significant with relation to dose are inverse in rate relative to the dose.

Overall the results suggest there are significant numbers of INDELs that accumulate in an Al^{3+} dose-dependent matter. The “AT” deletions and homopolymer insertions, especially the 2 bp INDELs, demonstrated an excess in both the P14 and *als3-3* genotypes. The 1 bp INDELs distribution was not linked to a pattern that explained their frequency.

Predicted Impact

In conducting analysis again using SNPeff, the predicted impact using all of the Al associated INDELs approximately 90% of the changes were either upstream, downstream or intergenic. Meaning that while these changes are occuring in the organism, given the location of the gene model where the INDELs were detected, these changes are not likely leading to an overall functional impact. Those INDELs that were located in an Exon were only between 1-3% with the rest of the INDELs being located in introns or in the UTRs of annotated genes. The full descriptive table below breaks down the predicted impact.

	P14 0.75 mM		als3-3 0.75 mM		P14 1.5 mM		als3-3 1.5 mM	
Type	Count	Percent	Count	Percent	Count	Percent	Count	Percent
DOWNSTREAM	704.5	35.70%	1020.75	36.57%	969.25	36.09%	1207.75	36.43%
EXON	56.5	2.86%	62.75	2.25%	53.25	1.98%	54.25	1.64%
INTERGENIC	285	14.44%	394.5	14.13%	386.25	14.38%	488.25	14.73%
INTRON	83.5	4.23%	118.75	4.25%	133.75	4.98%	162.25	4.89%
SPLICE_SITE_ACCEPTOR	2	0.10%	7	0.25%	3.5	0.13%	1.667	0.05%
SPLICE_SITE_DONOR	2	0.10%	1	0.04%	3.333	0.12%	1.667	0.05%
SPLICE_SITE_REGION	12	0.61%	15.666 66667	0.56%	13.667	0.51%	11.75	0.35%
TRANSCRIPT	2.667	0.14%	1.6666 66667	0.06%	1.5	0.06%	1	0.03%
UPSTREAM	779	39.47%	1102.75	39.51%	1056.25	39.33%	1313.5	39.62%
UTR_3_PRIME	29.25	1.48%	32.5	1.16%	33.75	1.26%	37	1.12%
UTR_5_PRIME	17	0.86%	34	1.22%	31	1.15%	35.75	1.08%

Table 26: Tabular Breakdown of Predicted Genomic Changes from Al Associated INDELs
This table is broken down into various categories of genomic changes based on the location of the INDEL in the gene model. Each genotype with its corresponding treatment is shown as an average of counts for direct comparison along with the corresponding percent of the total changes for the genotype and treatment concentration.

In evaluating these genomic changes the results suggest that while there is an increase in the counts that correlates to the increase in concentration of Al^{3+}, the percentages make an argument that the actual rate of where these changes are occurring in the gene model is not dose dependent. The most important takeaway of the predicted impact is the question of could these changes lead to mutations that could

affect the plant as a whole. One of the main factors for determining this would be the amount of genomic changes that affect the exons of the gene model. At a high of only 3% it seems very unlikely, additionally it appears that while other areas of the gene model are increasing in counts the those changes to exons remain about the same. While this will be discussed in the next section it is worth noting that this does correlate with previous research done in the Larsen lab. That these plants do accumulate mutations but they are not lethal suggesting that a possible repair mechanism could be mitigating the toxic effects.

Discussion

Experimental Setup

The original goal of this project was an attempt to identify genomic changes in the Arabidopsis plants in response to Al exposure and track them through generations. This was done through growing seedlings on gel soak media for 7 days and then allowing the plants to recover on a growth media before transferring the seedlings to soil. Then using pooled leaf tissue as a representative sample of the DNA of the organism, the genomic exposure and consequences to the Al^{3+} could then be extrapolated. By performing the experiment in this manner, the study would allow for not only a survey at the organismal level but also allow for a generational study. Since the plant can cope with the loss of the leaves but removing as essential as the root or the whole seedling would be lethal, and removes the possibility of a generational study if the tissue is harvested before the plant can go to seed. However this initial approach of only using a

pool of leaf tissue also led to confounding results and questioning how much Al^{3+} was actually present in the leaf tissue, additionally the previous set up included a recovery step. This recovery step could have led to the changes being observed as epigenetic changes instead of true genomic consequences. This also presented a statistical challenge as each replicate would only having a sample size of one seedling per generation would not provide enough statistical confidence to draw conclusions from the data, additionally there was no prior evidence to support how many generations would have to be completed to confirm any alternative hypothesis regarding Al^{3+} that was tested.

To address these issues with the initial setup, the experiment was revised to the setup previously stated where the whole seedling after being grown for 7 days on the gel soak media, however instead of recovering the plants on media and transferring them to soil, the seedling was collected and processed for DNA extraction directly from the gel soak media plate. This design choice of using the whole seedling was due to key information that was missing in the original setup, the primary source of Al^{3+} exposure for the plant comes from the soil via the root. However taking the whole seedling is more than just the root is justified in order to maintain reproducibility between genotypes and biological replicates. The growth difference between P14 and *asl3-3* is stark, *als3-3* itself grows a root that is barely visible to the naked eye. One criticism of this approach however is that the changes that are being analyzed are coming from the whole plant not just the root. Logically this could change the rate of the variants or other parts of the analysis such as percentage of changes in each sample, additionally in doing the analysis with the full seedling, denovo variants in the shoots could lead to additional

false positives. Overall the filtering process should mitigate these problems but further testing using just the roots to confirm this would be prudent. The suggestion to address this would be a method testing for the root specifically would be to dissect the root from the rest of the plant and use just that for the DNA extraction. However, specifically for *als3-3* to consistently get just the root, would require technical expertise that the lab currently does not possess. This would also need to be addressed in terms of experimental design how to be consistent between two different lines, which had dramatically different amounts of tissue.

Previous research with ATR and SOG1

Based on previous research done with ATAXIA TELANGIECTASIA MUTATED AND RAD3 RELATED (ATR) and SUPPRESSOR OF GAMMA RESPONSE 1 (SOG1) [26], and previous research by [16] found that in response to Al^{3+} exposure, Arabidopsis initiated a DNA Damage Response (DDR). Support for this claim comes from previous findings that Al^{3+} leads to Double Strand Breaks (DSBs) via findings of the COMET assay [17]. This finding only demonstrates that these double strand breaks are occurring but not how or by what molecular factors. The study that was performed did demonstrate that the presence Al^{3+} led to an increase in these DSBs.

In terms of the repair of DSBs there are two main factors that respond as part of the detection and response. The previously mentioned ATR and ATAXIA TELANGIECTASIA MUTATED (ATM) both these factors can play similar roles such as both can phosphorylate SOG1 leading to its activation [25,26]. In doing so a phosphorylated SOG1 furthers the DNA damage response (DDR) by leading to transcription of factors

required for the DDR [24], but ATR and ATM while similar respond to slightly different types of genomic damage, ATR typically responds at a more basal rate when occurrences like a replication fork stall occur leading to a DSB and persistent single stranded DNA [99,100] while ATM typically responds to more colossal damage and is typically seen as the organisms “overdrive” for DDR [101,61].

Since both are similar in that they respond to DSBs, but different to the level of their response and what level of genomic stress they respond to, it's important to understand which of these or both plays a role in the DDR when the plants are subjected to potential genomic stress from Al toxic environments. In order to determine this knockouts (KOs) of both of these genes *atr* and *atm* were crossed into the Al sensitive mutant (*als3-1)* at which time growth testing was done on both control (0.0 mM) and treated media (0.75 mM) to test if either could suppress the *als3-1* phenotype [26].

This had been done previously with *atr-4* as it was one of the first mutants identified by map based cloning to provide aluminum tolerance, and suppression of the *als3-1* hypersensitive phenotype [16,17]. However when this experiment was done with the doubles what was observed was that while *atr-4;als3-1* suppressed the *als3-1* hypersensitive phenotype, *atm-2;als3* did not [26,17]. These results demonstrate that ATR is the primary response factor from which the DDR is activated in response to Al^{3+} exposure. Under other conditions it does not preclude ATM from responding if conditions changed such as higher amounts Al^{3+}, or other stress factors were applied.

As the research shows ATR as the primary factor under which this DDR takes place as a result of Al^{3+} exposure, suggesting that some means of damage which ATR recognizes could be one of the primary sources of damage in which Al^{3+} is causing either

directly or indirectly. This could include the previously mentioned replication fork stalls and collapses [61],and possible conformational changes [101]. Meaning that Al^{3+} is potentially binding or changing DNA in some manner which the molecular machinery is detecting as damage, such as condensation of DNA [102]. There are three possibilities in which will be explored further.

First Al^{3+} could bind to the phosphate backbone of the DNA, which is very likely as the phosphates have a large negative charge which would attract the trivalent cation [103]. Al^{3+} could create a crosslink between nucleotides of either the same or opposite strand based on how other similarities in the physiological response of KOs to other crosslink agents [16]. Either of these options would be most likely to occur at the minor grooves (AT regions) of the DNA which would present the most likely target due to the availability both geometrically and the high amount of AT regions present in Arabidopsis [104].

Pipeline

While GATK [94] is not the only pipeline for determining genomic variants, it does employ more statistics to improve the sensitivity towards detecting rare variants. Other pipelines which could use tools such as Samtools with its variant caller VarScan2 [105] can be more user friendly and more straight forward on how the variants are identified. However for this study, especially some of these variants could be particularly rare in occurrence, GATK being more sensitive makes it the preferred choice. Additionally as previously mentioned that using an established pipeline helped to remove as much operator error as possible from the analysis.

SNPS vs INDELs

In terms of this study, the means in which the genomic consequences were identified were through the identification and classification of the genomic variants that were called from the sequencing samples. The two main types that were observed were SNPs and INDELs since there are many types of repair mechanisms that are at play in the DDR both of these could give clues in to what might be the most plausible mechanism of repair.

This is interesting in the context of what it means for the plant, most SNPs can be harmless to genomic integrity and stability of the plant. They can lead to possible substitution of amino acids in the generation of the proteins, but even then the Amino Acid (AA) code has redundancy built into translation. INDELs on the other hand can be very harmful if they fall in to a critical region such as an exon, where the loss or gain of nucleotides could change the protein through frameshift mutations, leading to gain, loss or just change of which amino acid is translated. The results of this study show there is a significant increase in 1-2 bp indels with Al treatment indicating that, as part of the DDR these indels are being created more often as the concentration of Al increases.

Al - Associated Changes

When dealing with these genomic variants it is important to note that in the analysis of these variants there were many categories for both the SNPs and the INDELs. After performing the GATK filtering (see analysis section), the binary table was used to identify those changes that were associated with Al exposure. Part of the

difficulty with performing this analysis is that studies have shown that Al^{3+} leads to DNA damage but not knowing how or what kind of damage forces the study to consider all possibilities.The binary table helps to clarify and categorize the different changes to allow the study to focus on just those changes related to the treatment.

However there were two categories which were confounding first were the High Frequency Changes (HFC), which represent those variants that only showed up in treated samples and were at the same genomic location with the same variant. The other was reported on briefly as the Al intermediate changes. Where the same variant was detected in the same genomic location in a control sample as one of the treated samples, but not in all the samples. However this variant was not present in both does levels of the treatment. While the HFC were left unexplored since they only represented a small portion of the overall changes, representing about six percent of the changes, the Al - intermediates being higher were explored. Based on the current sequencing results the best explanation would be that these changes were due to line changes (changes the differ from the reference genome) that had changed over time within each seed line and no longer showed up at a high enough rate in the population to be in both treatments. The possibility Also exists that these variants could have been changes that occured in treated samples to match that of the reference genome. Meaning that a variant in one line called as a A*T, could have become an A at this position which would have matched the reference genome and no longer would be called a variant.

AT Regions

Identified as part of this study is that predominantly the changes are increasing in frequency and are statistically significant are those deletions that called the variant as being adenine or thymine or both (AT) deletions. A previous study of cations from [46] noted that AT regions of DNA sequences seemed to be a primary target for cations to bind. However in these studies it was much easier to identify due to using cobalt as the cation which is much larger than other biological cations such as calcium (Ca^{+}) or Magnesium (Mg^{2+}), where as in this study using Al^{3+} in size is very close to Mg^{2+}.

This also makes a point about the possibility of Al toxicity causing DNA damage, if Al^{3+} is binding to the DNA backbone this potentially means that Al^{3+} is replacing Mg^{2+}, which is responsible in part for maintaining the structure and integrity of the DNA double helix [106]. AT regions being the location in the double helix of the minor groove which has the shortest location between each backbone would be the perfect location if Al^{3+} is creating cross links, or even if it's just the right location where it can match up its positive charges with the negative ones on the DNA.

Binding to the DNA can cause many out comes which would likely lead to a DDR this could include blocking transcription via replication fork stall due to blocking the RNA helicase [107,108]. Or larger effects such as possibly causing a conformational change or torsion changing the structure of the DNA from beta-form to psi-form DNA [109], which being condensed so could also lead to a replication fork stall. In *A. Thaliana* there is also a high AT content vs GC content which could play a part in why the AT regions of the DNA are significantly more likely to be the sites of indels. Additionally *A. thaliana* has a

65% AT content for its genome with GC content overall only reaching about 35% per chromosome [104].

Possible Genomic Consequences

AT content of *Arabidopsis* genome signifies a possible reason why these variants are of interest, even coding regions have approximately 55% AT composition [104]. There are also the possibility that is has to do with the structure of the DNA (discussed further on) as well as just numerical availability. There is also the possibility that these regions of AT are some of the strong points of the DNA [110] meaning that binding here could in fact change the conformation of the DNA leading to a more condensed form of DNA such as psi-DNA which could resemble methylated DNA or at least make those regions less accessible to transcription. This might also lead to the plant undergo epigenetic changes where it could only have the availability to access those genes not affected by the presence of the Al^{3+} cations (if they could not be repaired).

Overall however based on the predicted outcomes of the analysis (Tables: 17 and 26) it seems that these changes detected in the sequencing results and related to Al exposure seem to be mostly in non coding regions. Meaning that if Al^{3+} is causing changes its very rare to lead to mutation that would lead to an impact on the plants function. This should not down play the consequences however as the possible damage and response could lead to other molecular mishaps during normal cellular processes which could lead to worse consequences as discussed further with the DNA damage response.

DNA damage response

In terms of the current working model present previously under biologically relevant conditions, Al^{3+} makes its way through the plant vasculature and eventually finds its ways to the nucleus and DNA past all the other anionic sites that the plant has for it to bind to first. It then binds to the DNA [103] or in some way disrupts replication or transcription likely through a lesion. In doing so ATR detects either the DNA damage or a replication fork stall and halts the cell cycle while also activating SOG1 [26]. SOG1 can also if needed halt the cell cycle and initiate more of the downstream DDR. This is done by initiating transcription of key genes needed for DDR which can include genes like *BRCA1*, *PARP2* and more [26]. However this model while back by various studies and solid evidence does not mean its the only path for Al^{3+} detection and correction.

Types of repair that could involved

There are many different types of repair that the plant has to draw from the two most common are homologous recombination (HR) and Non Homologous End Joining (NHEJ) which has two forms its canonical (cNHEJ) form and its alternative (aNHEJ) form which is also referred to in some literature as Micro-homology Mediated End Joining (MMEJ). There are however many other forms of repair that include but are not limited Base Excision Repair (BER), Nucleotide Excision Repair (NER), and DNA Mismatch Repair (MMR).

HR is the plants prefered method of repair, which makes sense logically as its the repair method that uses the sister chromosome to which should have an almost if not identical region of DNA to use as a template of repair, where by ensuring almost error free repair [36]. This comes with the caveat that the sister strand would have to also possess the same region of DNA but could potentially have aberrations such SNPs for example. So in the scope of this analysis if this is the case then the results should show no increase in either SNPs or INDELs or potentially a rise in SNPs but not necessarily associated with the amount Al^{3+} the plant encounters. Which is what we see for the SNPs however the increase in INDELs suggests that in our system the plant either does not or can not use HR to try and remedy the problem created by Al^{3+}.

cNHEJ is one of the likely candidates for how these regions of DNA are being repaired, cNHEJ works by excising a segment of DNA that potentially contains a legion. In this case where Al^{3+} could be found to the DNA in some fashion as mentioned above.

After which it then recruits a ligation factor that then ligates the two sections of double stranded DNA back together, this form of repair is very prone to genomic changes, which would be seen as indels [111]. Which based on the data presented in this study would make sense that a form of repair just as cNHEJ could be the method by which the plant is using to repair these Al^{3+} sites.

MMEJ could also lead to loss of genomic information but creating indels, but does so in a more controlled way in which is requires two regions of homology on each side of the DNA to excised [111]. Hypothesizing that this could be the method by which the indels are created analysis was performed on the segments of DNA on either side of the indels, there was regions of homology discovered but no repeatable signature what discovered as to any sort signature for the binding site for this repair mechanism. Additionally as part of MMEJ There is a resectioning of the DNA which causes larger loss of genomic information. While this method of repair is possible is somewhat unlikely as the most significant INDELs that are seen as part of the treatment are only 1-2 bp in size.

Base Excision Repair (BER), could also be used to repair the damage especially when considering the oxidative nature of the possible damage. This could occur due the ROS response in which the base gets oxidised [112]. The oxidative damage as mentioned previously can lead to a lesion on the DNA, detection of this lesion can lead to activation of the DDR [113] and as previously mentioned lead to replication fork stalls. DNA glycosylases play a main role in BER, at least in humans, if the damage can not be repaired by these factors, the end result of BER being activated can lead to deletions [112,114]. ROS can also cause damage in the form of SSB which would also lead to the

activation of BER [115]. The end results lead to the possibility that due to the resectioning on the DNA insertion events could be occurring [116] or possibly a deletion event instead if the resectioning does not occur.

Nucleotide Excision Repair (NER),can also be used for the oxidative damage from the ROS response depending on where the damage occurs [112]. NER can also be activated due to interstrand crosslinks this could be occurred due to Al^{3+} [115]. Both of these types of damage could be possible other DNA damage such as treatments from chemicals like mitomycin C (MMC), or UV treatment [117]. These mutants that are sensitive to crosslinks have also shown sensitivity to Al^{3+} for example *rad17-1, lig6,* and *uvh-1* [16]. While there are some INDELs observed in the study that were detected could be large enough to represent events of genomic rearrangements the low rate of detection lead to these changes not being investigated further. But its possible as part of this process that INDELs of larger sizes could occur due to incorrect repair.

MisMatch Repair (MMR) may also play a role if the damage turns out to be due to alkylation of the DNA this could be alongside BER [41], It is also very prominent with replications errors which could be caused due to lesions or damage while replication is occurring [115]. Mismatch repair is reported as being high fidelity and should be preventing the any instance of INDELs [118]. While unlikely due to the data that was generated from the whole genome sequencing, MMR cannot be ruled out as potentially playing a role without further investigation.

Lastly of note there are many pathways such as fanconi anemia pathway [119] which use multiple different repair mechanisms. Rather than having one of these methods it could a combination of multiple different methods that has yet to be

determined. The possibility exists that due to the constant stress on the plant from the Al^{3+} in its environment that many of these processes could be working both in parallel or synergistically. With further research the hope is to shed light on this mystery.

Deus Ex Machina

One important note that requires further discussion is what is referred to in this section as the god or ghost in the machine. The reason being is that without adding evidence, to support that the Al^{3+} is in fact binding to the DNA it is hard to determine that the damage being caused is in fact due to Al^{3+} disrupting the molecular system of the plant, rather than the plant detecting the Al^{3+} and essentially over reacting and starting an unnecessary DDR that causes more damage than simply having the Al^{3+} in the system.

Support for this notion comes from growth tests of KOs of upstream regulators of the DDR, plants such as *atr-4* and *sog1-7*. These factors when knocked out allow the plants to grow as well or better than wild type on Al media, additionally double mutants of *atr-4;als3-1* and *sog1-7;als3-1* rescue the *als3-1* phenotype [16,26]. having these plant grow as well or better than wild type. In this study only P14 (wild type) and *als3-3* are used so unfortunately as yet there is no genomic data to also support this claim.

The reason this is interesting, is *atr-4* on its own could just signify that ATM which is fully functional in the plant is taking over the responsibility of activating SOG1, however also having *sog1-7* grow in the same manner helps to rule out that ATM merely fills in for ATR. Along this same vein, there seems as though some means of bypassing Al^{3+} exists, or that some other form of damage is occurring. This hypothesis comes not only by observing the growth of plants that have mutations in genes for DNA damage sensing factors but also by observing the molecular consequences of the Al sensitive mutant *als3-1*.

Even with the hypersensitive phenotype the plant is lead in to endo reduplication as part of some as of yet unexplained response to Al^{3+}. However endoreduplication which is a process by which the cell repeated makes more of its DNA in a repeated process that also leads to cell enlargement without cell division [26]. This means that the plant can in fact continue DNA replication despite the presence of Al^{3+}, additionally *sog1-7* and *atr-4* grown on Al media even as doubles with *als3-1* do not show this endoreduplication of cell enlargement. Which could suggest it is part of the downstream the DDR cascade that could be an overreaction to the presence of Al^{3+}.

There is a flip side to this argument, while DDR sensing KO mutants demonstrate Al tolerance, those plants with DDR factor KOs appear Al sensitive. This could signify two things the first being that these plants require these factors, so when they are knocked out the plant can not overcome and repair damage it would sustain normally outside of Al^{3+}, such as something like BRCA1 is also used to stabilize DNA during replication as part of having single stranded DNA (ssDNA) [120].

The second part of this flipside is that there could other factors that are initiating a DDR outside of ATR and SOG1 in response to Al^{3+} in a more controlled manner but still needs these factors, and overall becomes more sensitive due to other possible causes.

Other Factors

Plants are more complicated than they are often given credit for, in response to Al^{3+} and other stressors, plants possess numerous pathways to combat possible damage. Some of these methods include excluding Malate and Citrate to chelate cations such as Al^{3+} in acidic environments to prevent possible damage [121]. However there are other means once the Al^{3+} has entered in the plant system in which the plant can try to neutralize the threat of Al^{3+}. This includes the use of Reactive Oxygen Species (ROS), while this is a good means of preventing the Al^{3+} from further contaminating the plant's system, it comes with a draw back, too much of a good things also has consequences [40]. ROS is also a means by which DNA damage can occur, which also leads back to the ghost in the machine, it could be potentially the ROS response leading to DNA damage not the Al^{3+} directly.

Another part of the Al^{3+} response and a potential factor that could explain the phenomenon of the high frequency "A" and "T" change identified in the genomic analysis comes from the possibility that Al^{3+} could be leading to possible alkylation of the DNA, that in response would need a form of DNA repair such as BER, NER, or MMR to rectify the situation by removing the affected segments of DNA [41]. A programmed response of this nature could easily have some signature that is as of yet unknown but could be programmatically causing mutations in this fashion.

Materials and Methods

Seed Sterilization

Seeds were sterilized by first surface sterilizing with 200 uL 70% ethanol and 600 uL nuclease free water, which were vortexed and spun down at 8,000 RPM for 30 seconds. Then 600 uL of volume was removed and replaced with 600 uL of water for 4 times. After which 600 uL of volume was removed and 200 uL of bleach was added to further sterilize the seeds, the seeds were then vortexed and allowed to sit for not less than 5 minutes. After which 600 uL of nuclease free water is added and the seeds are vortexed, and spun down using previous settings. 600 uL of volume is removed and replaced with 600 uL of water for three times or until upon removing 600 uL the smell of bleach is no longer detectable, at which time as much liquid should be removed from the seeds as possible. Store the seeds at 2-8C for cold treatment and seed synchronization.

Growth

Approximately 15 seedlings of each genotype were surface sterilized with bleach and ethanol, cold treated at 4°C for not less than 3 days, this was done for each biological replicate. The plant material was then grown on Al gel soak plates [72] for 7 days, on both 0.0 mM and 1.5 mM Al treatments. The plates were composed of nutrient media 80 mL of 1 mM KNO3, 0.2 mM KH2PO4, 2 mM MgSO4, 0.25 mM (NH4) 2SO4, 1 mM Ca(NO3) 2, 1 mM CaSO4, 1 mM K2SO4, 1 mM MnSO4, 5 mM H3BO3, 0.05 mM CuSO4, 0.2 mM ZnSO4, 0.02 mM NaMoO4, 0.1 mM CaCl2, 0.001 mM CoCl2, 1%

sucrose, and 0.125% Gellan gum (Alfa Azer). The tissue was collected in 1.5 mL tubes that would be immediately flash frozen by liquid nitrogen (LN2), then stored at -80+/-10° C

DNA extraction

Seedlings are collected from media and placed individually in to 1.5 mL tubes, after which 250 uL Extraction buffer is added, which consists of DNA extraction buffer (0.35 M sorbitol, 0.1M Tris Base, 5 mM EDTA pH 7.5) with nucleic lysis solution (0.2 M Tris base, 0.05 M EDTA, 2M NaCl, 2% CTAB) and 5% Sarkosyl solution and sodium-meta-bisulfate. After which the tissue is disrupted with pestle and drill in the tube. 250 uL additional of DNA extraction buffer is added, and then incubated at 65 C for 1 hour. After the incubation 500 24:1 Chloroform: Isoamyl Alcohol, mixture is vortexed and then centrifuged at full speed (13,000 RPM) for 5 minutes. 400 uL of the aqueous layer is transferred to a new tube. 400 uL of ice cold isopropanol (IPA) is added to the new tube containing the 400 uL of the aqueous layer of the extraction. This tube is now vortexed and centrifuged for 15 minutes at max speed, then the supernatant is carefully poured off. 500 uL of 70% Ethanol (EtOH) is added to the tube and vortexed, then spun down at full speed for 5 minutes. After which the supernatant is poured off, ensuring not to disturb the pellet. The tube then needs to be be dried this can be done by allowing it to air dry but for consistency a speed vac was used buy running the speed vac with cooling for 30 minutes. The samples were then resuspended in 50 uL of nuclease free water to be sent out for fragmentation.

DNA Fragmentation

DNA samples were stored at -20C until shipped to University of California Irvine (UCI) Genomics Core, at which time the samples were shipped on dry ice to Irvine, California, where a covaris sonicator would fragment the DNA samples to 300-400 bp fragments and then had a 1-2 uL sample run on the bioanalyzer for confirmation of successful sheering. After the fragmentation was complete the samples were then shipped back on dry ice for library preparation.

DNA Library Preparation

Due to very small amounts of DNA collected from *als3-3*, the NEB ultra kit (Cat. #E7370S 24 reactions) was used to prepare the libraries. The fragmented DNA samples were brought up to a total volume of 55.5 uL and kit was followed with 2 minor deviations, the instructions for low DNA concentration were used in which the size selection took place after the PCR step of the library preparation, this is critical as the AMPure XP beads (Cat. #A63880) that are used for the clean up and especially for the size selection are based on concentration of the sample. The oligo adapters and primers used for the PCR came from BIOO Scientific, NEXTFlex DNA Barcodes (NOVA-514101).

Sequencing

Sequencing was performed at UCI the genomics core on an illumina HiSeq, three different barcodes were used one for each concentration of Al in the media, and were multiplexed by genotype. This would allow for greater depth to detect things past single nucleotide polymorphisms (SNPs) and allow for detection of INDELs and structural variations.

Analysis

The analysis of the genomic sequencing data was performed using a modified version of the SystemPipeR DNAseq pipeline. This was due to being run on a local machine rather than a compute cluster. The read mapping was done with BWA [92] after which the SAM (sequence alignment mapping) file was processed with picard tools [93] leading to the generation of the BAM (Binary alignment mapping) file. The GATK pipeline [94] was then followed for the generation and filtering of variant files using the default settings.

Modifications to the SystemPipeR include allowing the user to run the pipeline from the command line with dynamic trailing variables rather the static ones. Allowing for greater flexibility, the pipeline also checks to see if the time consuming steps have already been completed with the output files so as not to repeat them each time the analysis is run (such as the read mapping with BWA and generation of the SAM and BAM files). To conserve space after each main step is done in the analysis the files are compressed leading to improved storage.

Custom analysis outside of the pipeline provided by SystemPipeR was done to specify the analysis for Arabidopsis. After the variants were generated there was additional filtering done to remove any false positives for instance and genomic variants that had been been identified as part of the library construction or sequencing, but also to include things such as genomic segregation of alleles. This was done through a mendelian formulation, in which the variants were checked first if they were present in the same location and with the same change in one of the plants that been sequenced for the 0.0 mM Al plates, then if that variant showed up in 40% or more of the population it was removed.

The variants were then segregated into two different categories those variants which occurred only once and are considered denovo changes, and those that appeared multiple times but did not meet the criteria to be removed via the mendelian cutoff and are considered to be high frequency. Each set of variants were checked for any sort of pattern in relation to AL exposure compared to the control of 0.0 mM Al, with the *als3-3* and P14 being treated independently.

Statistical Testing

In order to determine significance for these the variants a global ANOVA was performed followed by and F-Test if there was determined significance to look at which factors were leading to the statistical significance. This testing was done with the functions native to R with guidance provided by Dr. Wenxui Ma.

Primers

Als3 wt 5’ : 5’-CAT GAA ACA GCT TCG AGA TGA C-3’

Als3 wt 3’ : 5’-AGC TGC TCC TAC CAT CAT GTT-3’

LBC Primer: 5’-TGG ACC GCT TGC TGC AAC TCT-3’

R Session Information

```
R version 3.2.1 (2015-06-18)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 16.04.4 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
 [1] grid      stats4    parallel  methods   stats     graphics
grDevices
 [8] utils     datasets  base

other attached packages:
```

```
 [1] reshape2_1.4.1            ggplot2_2.1.0
 [3] VennDiagram_1.6.17        futile.logger_1.4.1
 [5] rtracklayer_1.28.10       VariantAnnotation_1.14.13
 [7] ShortRead_1.26.0          BiocParallel_1.2.22
 [9] GenomicFeatures_1.20.6    AnnotationDbi_1.30.1
[11] Biobase_2.28.0            GenomicAlignments_1.4.2
[13] Rsamtools_1.20.5          Biostrings_2.36.4
[15] XVector_0.8.0             GenomicRanges_1.20.8
[17] GenomeInfoDb_1.4.3        IRanges_2.2.9
[19] S4Vectors_0.6.6           BiocGenerics_0.14.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.4          RColorBrewer_1.1-2   plyr_1.8.3
 [4] bitops_1.0-6         futile.options_1.0.0 tools_3.2.1
 [7] zlibbioc_1.14.0      digest_0.6.9         biomaRt_2.24.1
[10] RSQLite_1.0.0        gtable_0.2.0         lattice_0.20-33
[13] BSgenome_1.36.3      DBI_0.4              hwriter_1.3.2
[16] stringr_1.0.0        XML_3.98-1.4         latticeExtra_0.6-28
[19] magrittr_1.5         lambda.r_1.1.7       scales_0.4.0
[22] colorspace_1.2-6     labeling_0.3         stringi_1.0-1
[25] RCurl_1.95-4.8       munsell_0.4.3
```

Conclusions

Transcriptional Response to Al

Direct conclusions from the data

The RNA sequencing analysis chapter documented the 1% FDR used to employ a highly confident result set. However, these resulting gene lists contained primarily genes which were only predicted genes and with limited transcript support or with little detail of their annotated molecular functions. Further study is required with improved gene functional prediction pulling from othologous genes in other species or after further gene function curation. The improved functions might help confirmed if the expression based genes are linked to the Al response in plants. To that end since the overall molecular mechanisms leading to the stoppage of root growth are not understood these findings present a survey using the hypersensitive mutant als3-1. The best approach to helping this gap in understanding would like stem from testing already identified gene with known functions and performing additional testing to try to determine how these genes play in to the Al response in plants.

Additionally it is worth noting that there were many genes identified in this study that were *als3-1* specific. Potentially by lowering the FDR threshold to 5% the potential for greater overlaps between *als3-1* and Columbia-0 wild type (Col-0) exists. The down side of this approach is that the significant increase overall of gene targets would lead to increased need for stricter validation.

This would go past just performing technical validation of the results using methods such as qPCR and using in-vivo studies as under Al toxic conditions to show that that there would be an effect. While due to the complex nature of eukaryotic organisms, this could present its own challenge as the response could be very small and go unnoticed by the researcher.

Future experiments

In order to further pursue this subject, based on the confounding results of the original set of the genomics experiments it would be of interest to test for epigenetic changes due to exposure to Al. In testing this epigenetic change this could be another clue about key genes responsible for Al tolerance in plants. To do so would be relatively straightforward as many studies done with Al the plants are grown and placed into hydroponics. This could be one way of testing by looking at the growth before hydroponics and then after multiple timepoints of hydroponics with and without Al in the media. Using only relative growth rate, some initial conclusions can be drawn.

Similarly, in the set up using gel soak media, we saw confounding results when plants were rescued. The RNAseq analysis could be performed using plants that were treated and untreated with Al and a portion of each that was rescued and some that were not. This could go further and look at the overall study of the progenity to look for changes in the next generation when the experiment is repeated. This would require means of taking RNA samples before and after treatment to help fully identify the changes. Presenting a new comparison to look for trends in epigenetic changes that occur due to exposure. This would also be interesting to perform as generational study to see if these plants over time would become more tolerant to Al^{3+} as more generations

are examined. Ultimately this could provide valuable insight as to what genes are key to aluminum tolerance long term in plants. This would further prove targets to focus future research studies on in order to truly improve the understanding of the Al toxicity in plants.

Larger Scientific Impact

This experiment could ultimately be performed on any plant, Arabidopsis was chosen because it is much faster to grow and straight forward to extract RNA from. However, performing the analysis on another agriculturally relevant plant and performing comparative genomics could help identify the genes of interest that are related to Al response. While a direct comparison can not be be made overall gene models and similarities can be drawn especially for organisms that are not as well curated. In doing so could lead to increased speed of results of studies that could impact agriculture. Many target genes that have been found that relate to Al tolerance are not always cross species. If this approach works then it would allow for more in-depth research of this stress and hopefully the development of Al tolerant crops.

Genomic consequences

Direct Conclusions from the Data

One hypothesis is that direct or indirect exposure to Al^{3+} leads to a DDR and causes the formation of INDELs and loss of genomic integrity of the plant. These genomic lesions occur with some preference to AT sequences within INDELs. Further research is required to test the hypothesis of Al3+ can induce or promote thes liesions. I

postulate that evidence presented here indicates DNA damage repair is the primary source of these INDELs and potentially also contributes to the formation of SNPs. The variants observed in Al^{3+} treated samples demonstrated a bias toward “A” or “T” nucleotides. The mechanisms are not yet determined due in part to the difficulty of studying plants that are defective for factors related to DNA repair mechanisms, which are more sickly ans sensitive to stresses. Further study of how Al can gnerate a DDR should test if there is direct DNA interactions. Though this research provides some new evidence of the types of mutation biases in Al exposed plants, the mechanisms require additional genetic and molecular study.

Furthering of the model

My proposed working model for Al toxicity posits that when Al^{3+} permeates cell walls, despite primary defences to remove these ions, enters the vasculature and targets the nucleus. Through some means leading to a DNA lesion, and in turn a stalled replication fork and persistent single stranded DNA, ATR detects the anomoly and halts the cell cycle and activates the DNA damage response via phosphorylation of SOG1. If necessary upon activation SOG1 can halt the cell cycle and continue the DDR through transcription and regulation of various factors required for the DDR such as BRCA1 and PARP2. My research data indicates that in response to DDR, the sensing of Al^{3+} leads to excision of the DNA to either remove Al^{3+} along with its bound to DNA or the DNA interacting with Al^{3+} is cleaved leading to 1-2 bp INDEL or SNPs due to incomplete or inaccurate DNA repair. If the plant is unable to remove Al^{3+} alone when bound to DNA the INDEL mutatiosn are a conseqeunce of removal of the entire complex. This model

still lacks an explanation for why plants can cope in some cases with Al exposure and the mechanism of how endoreduplication is induced in hypersensitive mutant background.

I hypothesize that based on these results, further research on DNA damage factors will reveal how the damage being caused post Al exposure. This will help to determine if the damage is a result of direct exposure to Al^{3+} or if instead the damage is being caused by the DNA repair machinery trying to repair a problem that does not exist. The plant's molecular response to Al exposure is triggering ROS production and Alkylation and these processes in turn induce DNA damage. This could help explain the dose dependent effects observed in the data collected. It also offers an explanation in regaurds to the plateau in raw counts of genomic variants in 1.5 mM treated samples, and how these upper limits were only surpassed by using a hypersensitive mutant.

Future research

To further confirm the observations in this project and improve the understanding of the genomic consequences of Al toxicity, experiments repeating the Al exposure on *sog1-7* and *sog1-7;als3-1* genotypes and testing for relative changes in DNA mutation rates. TThis would improve confidence and quantification of the relative impact of Al sensitivty on DNA damage which has only been inferred from morphological observations in these genotypes. The examination of these mutants would also test if ATR and SOG1 are required to maintain the genomic integrity of the plant in response to the presence Al^{3+} or if they lead to overcompensation and greater genetic damage. This could be done in combination with additional analysis and refinement of this study this

includes but is not limited to doing further studies on to detect any significatn genomic hotspots using outside resources to help clarify further what is the real signature of damage caused in response to Al exposure.

Such outside resources would include comparison of the distribution of mutations found in wild accessions of Arabidopsis sequenced as part of the 1001 genome project [122] could further refine these results and improve our analysis by reomving variants that occur within natural popluations and isolating those changes that occur as a result of Al exposure. These resources could be used to highlight any additional false positives found in my data which might skew the results. Additionally with the knowledge gleaned from this work and methodology in analyzing the data, this experiment could be performed with other crop plants known to have genes that lead to Al tolerance or sensitivity to directly apply this study to real world agriculture. Pushing the boundaries further as to even change to other stresses or combination of stresses to determine the genomic impact on the plant.

If my experiments was to be repeated I would sequence the parent plant in addition and construct a draft reference genome that is more similar to the tested treatment samples. This would reduce counting of changes that are mutations that differ between the Col-0 that was sequenced for the TAIR10 genome and the starting accessions I used in my experiments, its possible that variant data from the 1001 genomes project [122] could also accomplish this goal. These data would be useful when comparing treated and untreated samples and improve the accuracy of the baseline mutation rates and potentially provide more confirnce in the observed trends.

Additional research for support

As a side experiement I have done experiments to create DNA crystals using the dickerson dodecamer in to test the hypothesis that Al^{3+} causes structural effects to DNA. Future work on this could be performed in collaboration with X-ray crystallography and structural biology experts to interpret the results and the possible biological significance. Application of NMR to explore the structural changes of DNA after Al exposure will provide additional biochemical perspective on mutations. A key question remains as to whether or not Al is directly binding to DNA and could further help understand how damage is leading to the generation of SNPs and INDEL and identify types of repair mechanism responding to Al exposure.

Hypothetical Experiments

My experiments were been conducted under sterile conditions, but I hypothesize that testing plants under realistic ecological conditions either by adding other stresses may reveal additional dynamics. One experiment could test the impact of Al in naturally occuring or contaminated acidic soil from places such as Belgium to test if mutation rates remain the same. Testing of further extremes in soil chemistry and Al content could help further understand the limits of plant tolerance. This research could also lead to the generation of crops with increased productivity in toxict soils which would support the gowing need for global food security.

Larger Scientific Impact

In the big picture, my research found detectable DNA damage resulting from Al^{3+} exposure in a dose dependent manner. This finding supports the use of genomic and transcriptomic approaches to explore mechanisms of DNA damage. My work suggests that further application genomic profiling of plants explosed to other stresses could further identify mutational biases and gene expression responses might point to whether similar damage and responses mechanisms are under similar molecular control. By studying the similarities and differences of plant responses to stresses I believe that the signature of each stress can be identified, and that those moleuclar response factors at the core of the general stress response can be identified. This would provide greater focus to different areas of study in plant stresses, with the potential prevent genomic damage to plants when stressed by the enviorment. By examining and contrasting the molecular consequences of stresses that induce DNA damage in similar ways, further testing could determine if Al treatment response resembles other stress responses such as those which cause overproduction of Reactive Oxygen Species or lead to faulty mismatch repair. Fundamental understanding of how Al damage is caused could improve chemical genomic screening approaches for factors that inhibit this damage or targeted breeding for Al resistance to develop Al tolerant genotypes for use in agriculture without reductions to yield or resilience of plants.

Bibliography

1. H.R. von Uexküll, Mutert E. Global extent, development and economic impact of acid soils. Plant Soil. 1995 Apr 1;171(1):1–15.

2. Ma JF, Hiradate S, Nomoto K, Iwashita T, Matsumoto H. Internal detoxification mechanism of Al in hydrangea (identification of Al form in the leaves). Plant Physiol. 1997;113(4):1033–9.

3. Negishi T, Oshima K, Hattori M, Kanai M, Mano S, Nishimura M, et al. Tonoplast- and plasma membrane-localized aquaporin-family transporters in blue hydrangea sepals of aluminum hyperaccumulating plant. PLoS One. 2012 Aug 29;7(8):e43189.

4. Kodama M, Tanabe Y, Nakayama M. Analyses of Coloration-related Components in Hydrangea Sepals Causing Color Variability According to Soil Conditions. The Horticulture Journal. 2016;85(4):372–9.

5. Miyasaka SC, Buta JG, Howell RK, Foy CD. Mechanism of aluminum tolerance in snapbeans : root exudation of citric Acid. Plant Physiol. 1991 Jul;96(3):737–43.

6. Delhaize E, Craig S, Beaton CD, Bennet RJ, Jagadish VC, Randall PJ. Aluminum tolerance in wheat (Triticum aestivum L.)(I. Uptake and distribution of aluminum in root apices). Plant Physiol. 1993;103(3):685–93.

7. Liu J, Magalhaes JV, Shaff J, Kochian LV. Aluminum-activated citrate and malate transporters from the MATE and ALMT families function independently to confer Arabidopsis aluminum tolerance. Plant J. 2009;57(3):389–99.

8. Ryan PR, Delhaize E, Randall PJ. Characterisation of Al-stimulated efflux of malate from the apices of Al-tolerant wheat roots. Planta. 1995 Mar 1;196(1):103–10.

9. Larsen PB, Cancel J, Rounds M, Ochoa V. Arabidopsis ALS1 encodes a root tip and stele localized half type ABC transporter required for root growth in an aluminum toxic environment. Planta. 2007 May;225(6):1447–58.

10. Kidd PS, Llugany M, Poschenrieder CH, Gunse B, Barcelo J. The role of root exudates in aluminium resistance and silicon-induced amelioration of aluminium toxicity in three varieties of maize (Zea mays L.). J Exp Bot. 2001;52(359):1339–52.

11. Barceló J, Poschenrieder C. Fast root growth responses, root exudates, and internal detoxification as clues to the mechanisms of aluminium toxicity and resistance: a review. Environ Exp Bot. 2002 Jul 1;48(1):75–92.

12. Brunner I, Sperisen C. Aluminum exclusion and aluminum tolerance in woody plants. Front Plant Sci. 2013 Jun 12;4:172.

13. Horst WJ, Wang Y, Eticha D. The role of the root apoplast in aluminium-induced inhibition of root elongation and in aluminium resistance of plants: a review. Ann Bot. 2010 Jul;106(1):185–97.

14. Larsen PB, Tai CY, Kochian LV, Howell SH. Arabidopsis mutants with increased sensitivity to aluminum. Plant Physiol. 1996 Mar;110(3):743–51.

15. Panda SK, Baluska F, Matsumoto H. Aluminum stress signaling in plants. Plant Signal Behav. 2009 Jul;4(7):592–7.

16. Nezames CD, Sjogren CA, Barajas JF, Larsen PB. The Arabidopsis Cell Cycle Checkpoint Regulators TANMEI/ALT2 and ATR Mediate the Active Process of Aluminum-Dependent Root Growth Inhibition. Plant Cell. 2012 Feb 1;24(2):608–21.

17. Rounds MA, Larsen PB. Aluminum-dependent root-growth inhibition in Arabidopsis results from AtATR-regulated cell-cycle arrest. Curr Biol. 2008 Oct 14;18(19):1495–500.

18. Heyman J, Cools T, Vandenbussche F, Heyndrickx KS, Van Leene J, Vercauteren I, et al. ERF115 controls root quiescent center cell division and stem cell replenishment. Science. 2013 Nov 15;342(6160):860–3.

19. Xia J, Yamaji N, Kasai T, Ma JF. Plasma membrane-localized transporter for aluminum in rice. Proc Natl Acad Sci U S A. 2010 Oct 26;107(43):18381–5.

20. Voragen AGJ, Coenen G-J, Verhoef RP, Schols HA. Pectin, a versatile polysaccharide present in plant cell walls. Struct Chem. 2009 Mar 13;20(2):263.

21. Luchi S, Koyama H, Iuchi A. Kitabayash i S, Kobayashi Y, et al (2007) Zinc finger protein STOP1 is critical for proton tolerance in Arabidopsis and co-regulates a key gene in aluminum tolerance. Proceedings of the National Academy of Sciences. 104(23):9900–5.

22. Fan W, Lou HQ, Yang JL, Zheng SJ. The roles of STOP1-like transcription factors in aluminum and proton tolerance. Plant Signal Behav. 2016;11(2):e1131371.

23. Kobayashi Y, Ohyama Y, Kobayashi Y, Ito H, Iuchi S, Fujita M, et al. STOP2 activates transcription of several genes for Al- and low pH-tolerance that are regulated by STOP1 in Arabidopsis. Mol Plant. 2014 Feb;7(2):311–22.

24. Yoshiyama K, Conklin PA, Huefner ND, Britt AB. Suppressor of gamma response 1 (SOG1) encodes a putative transcription factor governing multiple responses to DNA damage. Proceedings of the National Academy of Sciences. 2009;106(31):12843–8.

25. Yoshiyama KO, Kobayashi J, Ogita N, Ueda M, Kimura S, Maki H, et al. ATM□

mediated phosphorylation of SOG1 is essential for the DNA damage response in Arabidopsis. EMBO Rep. 2013 Sep 1;14(9):817–22.

26. Sjogren CA, Bolaris SC, Larsen PB. Aluminum-Dependent Terminal Differentiation of the Arabidopsis Root Tip Is Mediated through an ATR-, ALT2-, and SOG1-Regulated Transcriptional Response. Plant Cell. 2015 Sep;27(9):2501–15.

27. Yi M, Yi H, Li H, Wu L. Aluminum induces chromosome aberrations, micronuclei, and cell cycle dysfunction in root cells of Vicia faba. Environ Toxicol. 2010 Apr;25(2):124–9.

28. Collins AR. The comet assay for DNA damage and repair: principles, applications, and limitations. Mol Biotechnol. 2004 Mar;26(3):249–61.

29. Jia Q, den Dulk-Ras A, Shen H, Hooykaas PJJ, de Pater S. Poly (ADP-ribose) polymerases are involved in microhomology mediated back-up non-homologous end joining in Arabidopsis thaliana. Plant Mol Biol. 2013;82(4-5):339–51.

30. Krejci L, Altmannova V, Spirek M, Zhao X. Homologous recombination and its regulation. Nucleic Acids Res. 2012 Jul;40(13):5795–818.

31. Schuermann D, Molinier J, Fritsch O, Hohn B. The dual nature of homologous recombination in plants. Trends Genet. 2005 Mar;21(3):172–81.

32. Memisoglu A, Samson L. Base excision repair in yeast and mammals. Mutat Res. 2000 Jun 30;451(1-2):39–51.

33. Sarrasin B. Le plan d’action environnemental malgache de la genèse aux problèmes de mise en œuvre : une analyse sociopolitique de l'environnement. Revue Tiers Monde. 2007;n° 190(2):435–54.

34. Schärer OD. Nucleotide excision repair in eukaryotes. Cold Spring Harb Perspect Biol. 2013 Oct 1;5(10):a012609.

35. Spampinato CP, Gomez RL, Galles C, Lario LD. From bacteria to plants: a compendium of mismatch repair assays. Mutat Res. 2009 Sep;682(2-3):110–28.

36. Manova V, Gruszka D. DNA damage and repair in plants - from models to crops. Front Plant Sci. 2015 Oct 23;6:885.

37. Chao Q, Sullivan CD, Getz JM, Gleason KB, Sass PM, Nicolaides NC, et al. Rapid generation of plant traits via regulation of DNA mismatch repair. Plant Biotechnol J. 2005 Jul;3(4):399–407.

38. Buermeyer AB, Deschênes SM, Baker SM, Liskay RM. Mammalian DNA mismatch repair. Annu Rev Genet. 1999;33:533–64.

39. Hoffman PD, Leonard JM, Lindberg GE, Bollmann SR, Hays JB. Rapid accumulation

of mutations during seed-to-seed propagation of mismatch-repair-defective Arabidopsis. Genes Dev. 2004 Nov 1;18(21):2676–85.

40. Sharma P, Jha AB, Dubey RS, Pessarakli M. Reactive Oxygen Species, Oxidative Damage, and Antioxidative Defense Mechanism in Plants under Stressful Conditions. J Bot [Internet]. 2012 Apr 24 [cited 2017 Oct 24];2012. Available from: https://www.hindawi.com/journals/jb/2012/217037/

41. Fu D, Calvo JA, Samson LD. Balancing repair and tolerance of DNA damage caused by alkylating agents. Nat Rev Cancer. 2012 Jan 12;12(2):104–20.

42. Molinier J. Interchromatid and Interhomolog Recombination in Arabidopsis thaliana. THE PLANT CELL ONLINE. 2004 Feb 1;16(2):342–52.

43. Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem. 1998 Mar 6;273(10):5858–68.

44. Friesner JD, Liu B, Culligan K, Britt AB. Ionizing Radiation–dependent γ-H2AX Focus Formation Requires Ataxia Telangiectasia Mutated and Ataxia Telangiectasia Mutated and Rad3-related. MBoC. 2005 May 1;16(5):2566–76.

45. Yuan H-M, Liu W-C, Jin Y, Lu Y-T. Role of ROS and auxin in plant response to metal-mediated stress. Plant Signal Behav. 2013 Jul;8(7):e24671.

46. Arscott PG, Ma C, Wenner JR, Bloomfield VA. DNA condensation by cobalt hexaammine (III) in alcohol-water mixtures: dielectric constant and other solvent effects. Biopolymers. 1995 Sep;36(3):345–64.

47. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

48. GOrilla - a tool for identifying enriched GO terms [Internet]. [cited 2018 Dec 9]. Available from: http://cbl-gorilla.cs.technion.ac.il/

49. Poole RL. The TAIR Database. In: Edwards D, editor. Plant Bioinformatics. Totowa, NJ: Humana Press; 2007. p. 179–212.

50. Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. An “Electronic Fluorescent Pictograph” browser for exploring and analyzing large-scale biological data sets. PLoS One. 2007 Aug 8;2(8):e718.

51. Lee YJ, Szumlanski A, Nielsen E, Yang Z. Rho-GTPase-dependent filamentous actin dynamics coordinate vesicle targeting and exocytosis during tip growth. J Cell Biol. 2008 Jun 30;181(7):1155–68.

52. Gu Y, Li S, Lord EM, Yang Z. Members of a novel class of Arabidopsis Rho guanine nucleotide exchange factors control Rho GTPase-dependent polar growth. Plant Cell. 2006 Feb;18(2):366–81.

53. Yang Z. Cell Polarity Signaling in Arabidopsis. Annu Rev Cell Dev Biol. 2008 Oct 6;24(1):551–75.

54. Zhang Y, McCormick S. A distinct mechanism regulating a pollen-specific guanine nucleotide exchange factor for the small GTPase Rop in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2007 Nov 20;104(47):18830–5.

55. Yang Z, Fu Y. ROP/RAC GTPase signaling. Curr Opin Plant Biol. 2007 Oct;10(5):490–4.

56. Geldner N, Anders N, Wolters H, Keicher J, Kornberger W, Muller P, et al. The Arabidopsis GNOM ARF-GEF mediates endosomal recycling, auxin transport, and auxin-dependent plant growth. Cell. 2003 Jan 24;112(2):219–30.

57. Guo J, Wei J, Xu J, Sun M-X. Inducible knock-down of GNOM during root formation reveals tissue-specific response to auxin transport and its modulation of local auxin biosynthesis. J Exp Bot. 2014 Mar;65(4):1165–79.

58. Steinmann T, Geldner N, Grebe M, Mangold S, Jackson CL, Paris S, et al. Coordinated polar localization of auxin efflux carrier PIN1 by GNOM ARF GEF. Science. 1999 Oct 8;286(5438):316–8.

59. Uanschou C, Ronceret A, Von Harder M, De Muyt A, Vezon D, Pereira L, et al. Sufficient amounts of functional HOP2/MND1 complex promote interhomolog DNA repair but are dispensable for intersister DNA repair during meiosis in Arabidopsis. Plant Cell. 2013 Dec;25(12):4924–40.

60. Kerzendorfer C, Vignard J, Pedrosa-Harand A, Siwiec T, Akimcheva S, Jolivet S, et al. The Arabidopsis thaliana MND1 homologue plays a key role in meiotic homologous pairing, synapsis and recombination. J Cell Sci. 2006 Jun 15;119(Pt 12):2486–96.

61. Culligan KM, Robertson CE, Foreman J, Doerner P, Britt AB. ATR and ATM play both distinct and additive roles in response to ionizing radiation. Plant J. 2006 Dec;48(6):947–61.

62. Hong Z, Bednarek SY, Blumwald E, Hwang I, Jurgens G, Menzel D, et al. A unified nomenclature for Arabidopsis dynamin-related large GTPases based on homology and possible functions. Plant Mol Biol. 2003 Oct;53(3):261–5.

63. Miyagishima S-Y, Kuwayama H, Urushihara H, Nakanishi H. Evolutionary linkage between eukaryotic cytokinesis and chloroplast division by dynamin proteins. Proc Natl Acad Sci U S A. 2008 Sep 30;105(39):15202–7.

64. Ronceret A, Gadea-Vacas J, Guilleminot J, Lincker F, Delorme V, Lahmy S, et al. The first zygotic division in Arabidopsis requires de novo transcription of thymidylate kinase. Plant J. 2008 Mar;53(5):776–89.

65. Battaglia ME, Martin MV, Lechner L, Martínez-Noël GMA, Salerno GL. The riddle of mitochondrial alkaline/neutral invertases: A novel Arabidopsis isoform mainly present in reproductive tissues and involved in root ROS production. PLoS One. 2017 Sep 25;12(9):e0185286.

66. Xiang L, Le Roy K, Bolouri-Moghaddam M-R, Vanhaecke M, Lammens W, Rolland F, et al. Exploring the neutral invertase-oxidative stress defence connection in Arabidopsis thaliana. J Exp Bot. 2011 Jul;62(11):3849–62.

67. Schellmann S, Schnittger A, Kirik V, Wada T, Okada K, Beermann A, et al. TRIPTYCHON and CAPRICE mediate lateral inhibition during trichome and root hair patterning in Arabidopsis. EMBO J. 2002 Oct 1;21(19):5036–46.

68. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014 Jan;42(Database issue):D358–63.

69. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.

70. Werner T, Motyka V, Strnad M, Schmülling T. Regulation of plant growth by cytokinin. Proc Natl Acad Sci U S A. 2001 Aug 28;98(18):10487–92.

71. Yang Z-B, Liu G, Liu J, Zhang B, Meng W, Müller B, et al. Synergistic action of auxin and cytokinin mediates aluminum-induced root growth inhibition in Arabidopsis. EMBO Rep [Internet]. 2017 Jun 9; Available from: http://dx.doi.org/10.15252/embr.201643806

72. Larsen PB, Geisler MJB, Jones CA, Williams KM, Cancel JD. ALS3 encodes a phloem-localized ABC transporter-like protein that is required for aluminum tolerance in Arabidopsis. Plant J. 2005;41(3):353–63.

73. Berken A, Thomas C, Wittinghofer A. A new family of RhoGEFs activates the Rop molecular switch in plants. Nature. 2005 Aug 25;436(7054):1176–80.

74. Petukhova GV, Pezza RJ, Vanevski F, Ploquin M, Masson J-Y, Camerini-Otero RD. The Hop2 and Mnd1 proteins act in concert with Rad51 and Dmc1 in meiotic recombination. Nat Struct Mol Biol. 2005 May;12(5):449–53.

75. Vignard J, Siwiec T, Chelysheva L, Vrielynck N, Gonord F, Armstrong SJ, et al. The interplay of RecA-related proteins and the MND1-HOP2 complex during meiosis in Arabidopsis thaliana. PLoS Genet. 2007 Oct;3(10):1894–906.

76. H Backman TW, Girke T. systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics. 2016 Sep 20;17:388.

77. Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab cshl edu/fastx_toolkit. 2010;5.

78. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105–11.

79. Lutgens FK, Tarbuck EJ, Tasa DG. Essentials of geology. Pearson Higher Ed; 2014.

80. Macdonald TL, Martin RB. Aluminum ion in biological systems. Trends Biochem Sci. 1988 Jan;13(1):15–9.

81. Kochian LV. Cellular Mechanisms of Aluminum Toxicity and Resistance in Plants. Annu Rev Plant Physiol Plant Mol Biol. 1995;46(1):237–60.

82. Shirley BW, Hanley S, Goodman HM. Effects of ionizing radiation on a plant genome: analysis of two Arabidopsis transparent testa mutations. Plant Cell. 1992 Mar;4(3):333–47.

83. Hossain MA, Piyatida P, da Silva JAT, Fujita M. Molecular Mechanism of Heavy Metal Toxicity and Tolerance in Plants: Central Role of Glutathione in Detoxification of Reactive Oxygen Species and Methylglyoxal and in Heavy Metal Chelation. J Bot [Internet]. 2012 Apr 2 [cited 2018 Aug 15];2012. Available from: https://www.hindawi.com/journals/jb/2012/872875/abs/

84. Achary VMM, Jena S, Panda KK, Panda BB. Aluminium induced oxidative stress and DNA damage in root cells of Allium cepa L. Ecotoxicol Environ Saf. 2008 Jun;70(2):300–10.

85. Waterworth WM, Drury GE, Bray CM, West CE. Repairing breaks in the plant genome: the importance of keeping it together. New Phytol. 2011 Dec;192(4):805–22.

86. De Rybel B, Mähönen AP, Helariutta Y, Weijers D. Plant vascular development: from early specification to differentiation. Nat Rev Mol Cell Biol. 2016 Jan;17(1):30–40.

87. TAIR - About Arabidopsis [Internet]. [cited 2018 Jun 7]. Available from: https://www.arabidopsis.org/portals/education/growth.jsp

88. Koornneef M, Fransz P, de Jong H. Cytogenetic tools for Arabidopsis thaliana. Chromosome Res. 2003;11(3):183–94.

89. TAIR - Gene Annotation Data at TAIR [Internet]. [cited 2018 Jun 19]. Available from: https://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp

90. Larsen PB, Geisler MJB, Jones CA, Williams KM, Cancel JD. ALS3 encodes a phloem-localized ABC transporter-like protein that is required for aluminum tolerance in Arabidopsis: ALS3 is required for Al tolerance in Arabidopsis. Plant J. 2004 Dec 14;41(3):353–63.

91. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

92. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

93. Picard Tools - By Broad Institute [Internet]. [cited 2015 Oct 2]. Available from: http://broadinstitute.github.io/picard/

94. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

95. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014 Jul 15;30(14):2076–8.

96. Fisher 2x3 [Internet]. [cited 2018 Jun 13]. Available from: http://vassarstats.net/fisher2x3.html

97. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 2010 Jan 1;327(5961):92–4.

98. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012 Apr;6(2):80–92.

99. Zeman MK, Cimprich KA. Causes and consequences of replication stress. Nat Cell Biol. 2014 Jan;16(1):2–9.

100. Fulcher N, Sablowski R. Hypersensitivity to DNA damage in plant stem cell niches. Proc Natl Acad Sci U S A. 2009 Dec 8;106(49):20984–8.

101. Maréchal A, Zou L. DNA damage sensing by the ATM and ATR kinases. Cold Spring Harb Perspect Biol [Internet]. 2013 Sep;5(9). Available from: http://dx.doi.org/10.1101/cshperspect.a012716

102. Plum GE, Arscott PG, Bloomfield VA. Condensation of DNA by trivalent cations. 2. Effects of cation structure. Biopolymers. 1990;30(5-6):631–43.

103. Zhang R-Y, Liu Y, Pang D-W, Cai R-X, Qi Y-P. Spectroscopic and Voltammetric Study on the Binding of Aluminium(III) to DNA. Anal Sci. 2002;18(7):761–6.

104. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000 Dec 14;408(6814):796–815.

105. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr Protoc Bioinformatics. 2013 Dec;44:15.4.1–15.4.17.

106. Hartwig A. Role of magnesium in genomic stability. Mutat Res. 2001 Apr 18;475(1-2):113–21.

107. Hanas JS, Gunn CG. Inhibition of transcription factor IIIA-DNA interactions by xenobiotic metal ions. Nucleic Acids Res. 1996 Mar 1;24(5):924–30.

108. Recker J, Knoll A, Puchta H. The Arabidopsis thaliana homolog of the helicase RTEL1 plays multiple roles in preserving genome stability. Plant Cell. 2014 Dec;26(12):4889–902.

109. Antony T, Thomas T, Shirahata A, Sigal LH, Thomas TJ. Selectivity of spermine homologs on triplex DNA stabilization. Antisense Nucleic Acid Drug Dev. 1999 Apr;9(2):221–31.

110. Bacolla A, Tainer JA, Vasquez KM, Cooper DN. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res. 2016 Jul 8;44(12):5673–88.

111. Bétermier M, Bertrand P, Lopez BS. Is non-homologous end-joining really an inherently error-prone process? PLoS Genet. 2014 Jan;10(1):e1004086.

112. Cooke MS, Evans MD, Dizdaroglu M, Lunec J. Oxidative DNA damage: mechanisms, mutation, and disease. FASEB J. 2003 Jul;17(10):1195–214.

113. Ciccia A, Elledge SJ. The DNA damage response: making it safe to play with knives. Mol Cell. 2010 Oct 22;40(2):179–204.

114. Wyatt MD, Allan JM, Lau AY, Ellenberger TE, Samson LD. 3-methyladenine DNA glycosylases: structure, function, and biological importance. Bioessays. 1999 Aug;21(8):668–76.

115. Maynard S, Schurman SH, Harboe C, de Souza-Pinto NC, Bohr VA. Base excision repair of oxidative DNA damage and association with cancer and aging. Carcinogenesis. 2009 Jan;30(1):2–10.

116. Beard WA, Prasad R, Wilson SH. Activities and Mechanism of DNA Polymerase β. In: Methods in Enzymology. Academic Press; 2006. p. 91–107.

117. Kondo N, Takahashi A, Ono K, Ohnishi T. DNA damage induced by alkylating agents and repair pathways. J Nucleic Acids. 2010 Nov 21;2010:543531.

118. Culligan KM, Hays JB. DNA mismatch repair in plants. An Arabidopsis thaliana gene that predicts a protein belonging to the MSH2 subfamily of eukaryotic MutS homologs. Plant Physiol. 1997 Oct;115(2):833–9.

119. Moldovan G-L, D'Andrea AD. How the fanconi anemia pathway guards the genome. Annu Rev Genet. 2009;43:223–49.

120. Aparicio T, Baer R, Gautier J. DNA double-strand break repair pathway choice and cancer. DNA Repair . 2014 Jul;19:169–75.

121. Wang Y, Cai Y, Cao Y, Liu J. Aluminum-activated root malate and citrate exudation is independent of NIP1;2-facilitated root-cell-wall aluminum removal in Arabidopsis. Plant Signal Behav. 2018 Jan 2;13(1):e1422469.

122. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011 Aug 28;43(10):956–63.